Exploiting Linguistically-Enriched Models for Phrase-Based Statistical Machine Translation

Noemie Guthmann

Master of Science

Speech and Language Processing

School of Philosophy, Psychology and Language Sciences

University of Edinburgh

2006

Abstract

This thesis presents the design and implementation of linguistically-informed models for statistical phrase-based machine translation. Using Koehn's Pharaoh (2004), a state-of-the-art SMT system, and Moses (Hoang, 2006), a variant of the former which supports factored translation models, we have investigated two approaches: Combined Feature Models and Factored Models. While Combined Feature Models make use of concatenations of linguistic features to enrich their models, Factored Models view a token as a vector of factors, enabling to build relatively independent models for each factor. In the context of machine translation, both models were expected to enrich the existing surface word model with additional linguistic information.

The research undertaken focused on finding ways to improve output translation quality for English-to-French and French-to-English translations from various standpoints. A better general readability and understandability of a generated document should be achieved mainly by ensuring the text fluency in the target language (syntactic correctness), its adequacy (use of adequate terminology) and its fidelity (semantic adequacy). These main goals were addressed by first of all analyzing the Pharaoh's current performance, and understanding language-specific and model-related problems encountered. Several experiments were then performed using our two approaches, and their results were compared.

Despite a few noted improvements in some of the linguistic issues discussed, notably fixed expression translation and part-of-speech ambiguity, major problems involving complex syntactic structures in the source language still posed a hard challenge to the approach of linguistically augmenting phrase-based statistical machine translation.

Acknowledgments

I would like to thank first of all my supervisors, Dr Philipp Koehn and Dr Mirella Lapata, for their assistance, patience and motivation in this project; you helped me exploit the potential of this thesis, and I am very grateful for it.

Academically, I would also like to thank Phd students Hieu Hoang and Abishek Arun, whose willingness to help was very much appreciated.

I thank also my friends and family here and overseas for their constant support: to my friends and study partners, Sharon Givon, “stuck in the middle with you”, from whom I learned so much academically and personally, and Hamutal Meridor, who has been willing to help whenever needed. To my family, who has lived every minute of this year with me and stopped breathing at stressful times. To Natan Devir, ever there for me. And to Iddo Greental, whom I thank for his comfort and precious love. All of you, this thesis could not have been done without you.

Finally, I would like to acknowledge the generous University of Edinburgh UK/EU Masters scholarship, which made the entire experience possible.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Noemie Guthmann)

Table of Contents

Chapter I: Introduction

The data revolution has greatly influenced the field of Machine Translation: researchers have tried to outgrow traditional rule-based methods by using statistical models, which could learn linguistic patterns automatically from data. Despite some encouraging results on the output translation grammaticality, adequacy and comprehensibility, much improvement is still to be achieved. Recent research in machine translation is thus seeking new ways to significantly improve the quality of the output, by addressing the different parts of the translation process. The present thesis occurred in the framework of a shift in direction for statistical models in MT, that tends to turn towards the integration of linguistic knowledge as a potential breakthrough in translation performance.

1.1. Phrase-based machine translation

Statistical machine translation systems are based on probabilistic models automatically induced from corpora. These systems do not include a separate language generation module, as many non-statistical machine translation systems do. Instead, the principle on which they rely to generate grammatical sentences in the target language is a calculation of the cheapest cost for the best combination of hypotheses out of a space of possibilities. Classic statistical machine translation (SMT) systems implement the noisy channel model: given a sentence in the source language f, we try to choose the translation in language e that maximises p(e|f). According to Bayes rule, this can be rewritten as:

$$\arg\max_e \; p(e \mid f) = \arg\max_e \; p(f \mid e)\, p(e)$$

p(e) is materialised with a language model – typically, a smoothed n-gram language model in the target language – and p(f|e) with a translation model – a model induced from parallel corpora – aligned documents which are the translation of each other. Several different methods have been used to implement the translation model, and additional models such as fertility and reordering models have also been employed, as in among the first translation schemes proposed by the IBM Models 1 through 5 in the late 1980's (Brown et al, 1993) (see also Knight and Marcu, 2004). Finally, the decoder is an algorithm that calculates the most probable translation out of several possibilities, derived from the models at hand.

Pharaoh (Koehn, 2004) is a state-of-the-art phrase-based beam search decoder for statistical machine translation. It relies upon several models, including the language and translation models described above, and a decoding algorithm. The translation model used by Pharaoh is trained from parallel corpora using word alignment methods, and includes a probability distribution over phrase pairs (rather than just single words) of source and target languages. Additional models (a distortion model and word penalty) are included in the best translation calculation, which is searched for by beam-search decoding.

Phrase-based models have proved to significantly outperform word-based translation (Koehn et al, 2003). However, phrase-based systems performance leaves space for much improvement from the standpoint of their outputs' grammaticality and understandability. Indeed, the phrases that Pharaoh currently uses for translation are not syntactically motivated, which may cause ungrammatical output. Furthermore, the models in question tend to perform much better when translating to morphologically simpler languages. Phrases used for translation are exploited at the lexical level, and so the morphological and syntactic complexity of the source text is unutilized, which may create difficulties for morphologically-richer languages such as French.

In this thesis, we argue that richer structural information may enhance current SMT models and thus improve translation performance.

1.2. Using linguistically-informed models for SMT: a review of what has been done

Ongoing research aimed at improving these available SMT methods take the approach of augmenting the models at hand with linguistic information, from word-level information (morphology) to syntax. Linguistic information can be supplemented in several ways and using various methodologies.

One area of intensive research is automatic word alignment, on which translation models are based. For instance, Corston-Oliver and Gamon (2004) have explored the impact of word morphology on word alignment of a parallel corpus. They have achieved better word alignment by reducing vocabulary size through stemming of the English-German parallel corpus, thus aligning related word forms. In Niessen and Ney (2001), hierarchical lexicon models enabled the word alignment to interpolate counts based on different combined linguistic features representing words. Giménez and Màrquez (2005) have also achieved

better translation performance with Pharaoh by feeding the alignment algorithm with combined linguistic representations of words.

Researchers have also tried to enrich the translation model with linguistic features: for example, (Och & Ney, 2002) have used an alternative statistical method to Bayesian statistics in the translation model, called Maximum Entropy, to enable the definition of linguistic features in the translation models. The main difficulty with this approach seemed to be the selection of features to use. A particularly interesting approach, which has been the subject to ongoing research in the field, and first experimented at John Hopkins 2006 Summer Workshop on SMT, is to improve translation models using factored models. These models relate to lexical items as vectors of factors designating linguistic information (e.g. stem, morphological inflection, part of speech).

Better language models can also enhance statistical machine translation performance: experiments with factored language models in automatic translation have been pursued by Yang and Kirchhoff (2005). Yang and Kirchhoff (2006) also developed another method that has found gain in using linguistic information: they proposed phrase-based backoff models for unknown words using hierarchical morphological abstractions at the word and phrase level: this significantly improved the quality of translations by reducing the number of unknown words that were previously simply inserted in their original form.

Finally, we shall discuss here syntax-based SMT, which has been intensively investigated over the past few years. The advantage of such an approach is that it takes into account language-specific word order and deals with long-distance constraints, as opposed to previously mentioned statistical models. Yamada and Knight's syntax-based translation model (2001) takes flattened syntactic trees as input obtained using Collin's (1997) statistical parser, and maps them to a string sentence in the target language. The trees are then fed into the system: children nodes in the tree are reordered, extra words inserted at each node. Finally, leaf words are translated. Yamada and Knight's system was later extended to support syntactically motivated phrasal translation (rather than just word-based), and in subsequent work, Charniak, Yamada and Knight (2003) added a syntax-based language model to select the most probable parse. This last model has proved to outperform IBM's Model 4 translation quality. Finally, Chiang's (2005) work on hierarchical phrase-based models has proposed to incorporate a synchronous context-free grammar induced from un-annotated parallel corpora to a phrase-based model. Chiang reports a BLEU score improvement of 7.5% for Mandarin-to-English translation compared to Koehn's Pharaoh.

1.3. Our approach: using Combined Linguistic Features and Factored Translation Models in SMT

In this thesis, we wished to deal with syntactic and lexical problems encountered in the task of Statistical Machine Translation, while building augmented translation models for French-to-English and English-to-French. These models would exploit the rich linguistic information inherent in the source and target texts of the parallel corpus. We suggested two different methods. The first one was to enrich the translation models used by the Pharaoh decoder with combined linguistic features, as had previously been done by Giménez and Màrquez (2005) on English-Spanish translation models. The second was to incorporate factored translation models to current phrase-based methodology, using Moses (Hoang, 2006), a variant to Pharaoh's phrase-based decoder which implements factored models. By training models on different levels of linguistic representations of words, we wished to integrate this additional information in the model to hopefully direct it to choose more correct translations from the viewpoint of their grammaticality and understandability. Both implemented methods thus required linguistic annotation of the parallel corpora, which was performed with various existing NLP tools for each language in question. From there on, the methodology differed as to how these augmented corpora were utilized. We shall explain both.

To build combined feature models, each token had to be regarded not just as a word form anymore, but as the representation of that word at various linguistic levels. The features (or tags) for a word were thus concatenated: for example, lexical and morphological information could provide the word's base form and inflectional information, while the part-of-speech would give some information on the word's class and its syntactic role in the sentence. So, for instance, the word "results" in the expression "this results in" could be represented as "results_VBZ", as opposed to "results_NNS", indicating that in this context, this word acts as a verb in the third person singular and conjugated at the present tense, rather than as a common noun in its plural form. This approach was thus expected to fit well the phrase-based approach for SMT, as it takes into account the direct environment of a word be to translated. We hoped to help improve translation performance in three ways: firstly by building a translation model where linguistic representations of words would enforce correct translations of the source text in the context of certain word sequences; secondly by building language models where those features would give a bigger weight to the correct use of words and word sequences in the target language; and thirdly, by exploiting the linguistic

information in the test set (or source text to be translated). Indeed, if we have worked in the same line of thought as in Giménez and Màrquez (2005) when designing our experiments, we departed from them in the sense that we went beyond just dealing with the word alignment. In addition to feeding the alignment algorithm with combined linguistic features, we also annotated linguistically the test data in order to help disambiguate it and make use of the linguistic information from enriched translation models. We thus decided not to strip our translation tables off from linguistic information (as had previously been done by Giménez and Màrquez), but rather to translate test data represented as linguistic features combinations. The following illustrates how we expect to decode a phrase with our models:

$$(je|\text{PRO})(vous|\text{PRO})(achète|\text{VER(pres)}) \rightarrow (i|\text{PRP})(buy|\text{VBP})(you|\text{PRP})$$

In our second approach, we aimed at enriching the current models by using factored translation models to represent the additional information inherent in the training corpus. Moses is a decoder under development which is based on Pharaoh, but allows the use of multiple factors in the translation models. In a factored translation model, each word is represented as a vector of linguistic factors. In the context of machine translation, these models are expected to enrich the existing surface word translation model with additional linguistic feature models. The major advantage of factored models is that they enable the training of linguistically-informed factors independently from each other, thus avoiding the effect of sparse data, but allowing to keep track and manipulating in different ways a word's multiple factors. The factored models being currently under development and in constant adaptation to the task of Machine Translation, we limited our experimentation to the impact of factored models over the target language, or to what is called the generation step. Factored models as implemented in Moses are expected to first *translate* phrases of factors from source to target language, and then limit on the possible factor sequences by applying *generation* from translated target factors to other target factors. In the framework of this research, we translated surface words only, and then used generation tables and separately trained language models on factors for the target language, to pose a linguistic constraint on the decoder's search for the best translation. Factors were selected from linguistically-relevant properties that were expected to bear influence on translation quality. The inventory of factors used included word lemma, inflected form and part of speech. The following illustrates how French to English translation should be performed on phrases with these models using the translation and generation steps (Hoang, 2006):

$$(je)(vous)(achète) \rightarrow (i)(buy)(you)$$

Figure 1 : Translation Step

$$\begin{pmatrix} i \\ \downarrow \\ PRO \end{pmatrix}\begin{pmatrix} buy \\ \downarrow \\ VB \end{pmatrix}\begin{pmatrix} you \\ \downarrow \\ PRO \end{pmatrix}$$

Figure 2 : Generation Step

Finally, we proposed to verify our models' robustness against training on limited data. Indeed, acquiring parallel corpora is not an easy task, especially for languages other than the main European languages for which little linguistic data and corpora have been collected to date. Enabling to train on limited data being one incentive for developing linguistically-informed statistical models, we thus created baselines for each model trained on a portion of the parallel corpus, as well as on the whole corpus, to have an idea of the impact of corpus size on our models.

1.4. Work structure

Chapter one, this current chapter, has been an introduction to state-of-the-art statistical machine translation, and to our suggested approaches to improving it: we have given an overview on SMT and what it currently does and does not achieve. We have presented the different research efforts of the past few years aimed at integrating linguistic knowledge in the statistical models used by machine translation systems. Finally, we have described the two approaches our research has implemented to augment these models with linguistic information.

Chapter two establishes the baselines using the state-of-the-art statistical machine translation system, Pharaoh, on French-English translations in both translation directions. We propose evaluation methods and apply them to our baselines to define the problems we wish to deal with.

Chapter three defines our methodology. This includes setting our research aims and building hypotheses according to findings from the previous chapter, in terms of the linguistic features that should be used. We then suggest how these hypotheses should be implemented

with the two approaches proposed. We explain the tools and methods chosen for data preparation, present how the different models were trained and tested.

Chapter four presents the results of the experimentation, checks them against the baseline results and against the initial hypotheses, and suggests some explanations to the various findings.

The fifth and last chapter concludes this thesis and proposes some directions for future work.

Chapter II: Baseline Setting

First and foremost, we wished to enquire Pharaoh's current performance, based on French-to-English and English-to-French translations. We then looked for various ways of improving this performance using different models: on the one hand, combined linguistic feature models, and on the other, factored models.

Baseline models were established for English-to-French and French-to-English translation: Pharaoh's phrase-based translation model was trained on different sizes of the French-English Europarl parallel corpus (first on 50,000 sentences, then on the whole corpus) to investigate the effect of scarce data on our models. These models were studied through the comprehensive evaluation of their output by using several methods: automatic evaluation methods included the BLEU metric, as well as a more fine-grained automatic evaluation method; manual evaluation was performed on a sample of 150 sentences from the outputs of baselines trained on the smaller portion of the training corpus. The baseline models' advantages and limits were investigated in order to define a new line of research.

2.1. Building the baselines

Eight baselines were built for the various experiments. They used only surface words and no additional features. 4 models were built for comparison with the combined features models: 2 models – for French-to-English and English-to-French translations – were trained on a section of the Europarl parallel corpus (50,000 sentences), and 2 other models were trained on the whole corpus (composed of approximately 700,000 sentences). 4 additional baseline models (2 on the limited corpus size, the other 2 on the whole corpus) were trained without a model for lexicalized reordering, for comparison with the factored models which did not support such models.

Firstly, the parallel corpora were fed into the GIZA++ word alignment algorithm (Och and Ney, 2000) and some additional heuristics to extract phrase alignment. Minimum Error Rate Training (Och, 2002) was then applied to refine these models, using a tuning French-English parallel corpus of between 300 to 1000 sentences. Translation tables were filtered to adjust to the test data, a section of 2000 sentences of the same domain as the training data, but which were not part of this data. Finally, each baseline model was tested using either Pharaoh

(for the first 4) or Moses (for the last 4) decoders. Here, Moses was configured to decode in a very similar way as Pharaoh (i.e. without using factored models).

2.1.1. Lexical phrase-based translation with Pharaoh

In Pharaoh, the calculation of the best translation is mainly based upon a translation model and a language model. These models are implemented with a phrase translation table, where translation probabilities for phrase pairs are stored, and a smoothed n-gram language model of the target language. In addition, a reordering model and a word penalty model are computed.

$$p(e \mid f) = p_{\phi}(f \mid e)^{\lambda_{\phi}} \times p_{LM}(e)^{\lambda_{LM}} \times p_{D}(e, f)^{\lambda_{D}} \times \omega^{length(e)\lambda_{W(e)}}$$

As can be seen above, these models are weighted, and their product enable the system to rank translation hypotheses according to their probability of representing a correct translation in the target language. The algorithm that performs that calculation is called the decoder: it expands a space of hypotheses based on the probabilities from the models, and performs a search through this space for the best hypotheses. This search is maximized using hypothesis recombination, but also pruning methods such as future cost estimation.

2.1.1.1. Parallel corpus

The models used in a statistical machine translation system are trained on parallel corpora: a parallel corpus is composed of sentenced aligned (one sentence per line) documents which are the human translations of one another: each sentence on one side of the corpus is the translation of that same sentence on the other side.

The parallel corpus used for these experiments was the French-English Europarl corpus, which includes for each language around 20 million words, and 700,000 sentences. Europarl is collected from the Proceedings of the European Parliament, and has been pre-processed for use with SMT systems (Koehn, 2005) including sentence splitting and tokenization, as well as lowercasing (to avoid training separate models on uppercase and lowercase words). The text was meant to be outspoken in the European Parliament meeting sessions, and so it was written in a high-level style, sometimes in direct speech – when the speaker addressed himself to its audience, sometimes in indirect speech – when someone else's sayings were being reported. Often, a narrative style was used as external facts were being brought forward by the speaker (such as argumentation for his/her speech) or by the transcriber (such as the time and place of the session). Most of the issues discussed had to do with European internal policies and laws.

This mixture of direct, indirect and narrative speech, the formal nature of the circumstance for which these speeches were prepared to be performed, and finally, the relative seriousness of the topics themselves, all had great influence over the style and domain of the corpus, which can be seen as high-level formal speech. This had obvious implications over certain grammatical choices in both languages, such as avoiding the use of the colloquial second person singular in French, as well as the Imperative tense in English for the second person singular.

2.1.1.2. Building the translation models

As mentioned earlier, the translation model in Pharaoh is composed of a translation table and a distortion (or reordering) model. These are automatically induced from a parallel corpus.
Phrase translation tables represent phrases in the source language and their possible translations into the target language, graded with probabilities as automatically learned from the parallel corpus. In Pharaoh, these phrases are not linguistically motivated as these proved to decrease the performance; a phrase is thus a sequence of n word tokens (starting from n=1), including punctuation.

The algorithm used for word alignment is the EM (Expectation-Maximization) algorithm proposed in GIZA++, a freely available implementation of the IBM models (Brown et al., 1993). This algorithm aligns tokens in sentence pairs extracted from the parallel corpus and finds the most likely word alignment by iterative search. Pharaoh makes use of bidirectional runs of GIZA++: this is because one run of the algorithm can only generate one-to-many translation, from target to source language. The heuristics then used to extract phrase alignment are described in (Och et al, 1999); in brief, word alignment is extracted by intersecting these two alignments, and phrase pairs are then collected that are consistent with the word alignment. The translation table, which represents the probability of source (f) language phrases translation into target (e) language phrases (or $\phi(e \mid f)$) is then built by computing a probability distribution by relative frequency over these phrase pairs:

$$\phi(\bar{f} \mid \bar{e}) = \frac{count(\bar{f}, \bar{e})}{\sum_{\bar{f}} count(\bar{f}, \bar{e})}$$

It shall be noted that no smoothing is performed on the translation table, relegating the sparse data problem to lexical weighting. Reordering is modelled by a relative distortion probability distribution over the sentence pairs.

2.1.1.3. Building a language model

A language model is a statistical model of which parameters are learned from corpora: word sequences (or n-gram) probabilities are estimated by computing their relative frequency in the corpus. The language model toolkit we used in our experiments was the freely available SRILM toolkit (Stolcke, 2002). We created trigram language models for our baselines, learnt from the Europarl corpus in the target language: indeed, it is important that the language model is of the same domain as the translation model and the test set. Discounting and smoothing methods (interpolation and Kneser-Ney smoothing) were used to deal with the problem of unseen events.

2.1.1.4. Minimum Error Rate Training

Minimum Error Rate Training, or MERT (Och, 2002), optimizes translation quality by setting the model weight parameters. This is done by taking a held-out section of the parallel corpus, running the decoder with its current translation model on the source language text, and then automatically evaluating the output's translation quality by comparing it to real translation (using automatic methods such as BLEU and word error rate). The weights attributed to the current models are then adjusted accordingly, and the process is iterated until convergence. To run MERT, we thus define starting values and ranges for the distortion, translation and language models parameters, the development set and the models in use.

2.1.1.5. Filtering the phrase table

Filtering the phrase table according to the test set we intend to use enables to tune the decoding process for memory usage (Koehn, 2004). Indeed, by limiting the phrase table to phrases that appear in the test data and their potential translations, we avoid loading the entire phrase table.

2.1.1.6. Beam search decoding

Pharaoh's decoder can translate files one sentence per line in the source language. To translate a sentence, the decoder generates a first hypothesis, or partial translation of a phrase in the input. Then, another hypothesis is generated, based on the previous: the decoder keeps a stack of the best partial translations until now. The notion of "best", or "low cost" is equivalent to "most probable", where probabilities for a hypothesis are the product of probabilities given by the models discussed above.

The decoder uses several methods to limit the search space, including recombination of hypotheses, which is risk-free, and beam search, which risks the pruning of good translation hypotheses. This search algorithm estimates hypothesis cost based on both the future cost (a possibly pre-computed calculation of the part of the sentence which has not yet been decoded, including the language model and translation model factors) and the cost so far, and prunes out more costly hypotheses to only expand those that are likely to succeed. The future cost calculation does not however take into consideration the reordering cost; also, it only gives an estimate of the language model cost. It is thus prone to error. Eventually, the best scoring final translation is outputted.

The decoder reads from a configuration file which indicates where the translation models are located, as well as the different weights to these models. The beam size can be defined with a threshold or by histogram pruning: we used the default threshold, which cuts off probabilities that are less than 0.00001. The maximum stack size for holding hypotheses was set to 100. We also defined a standard distortion limit (maximum distance between two input phrases to two neighbouring output phrases – see (Koehn, 2004)) of 4, as well as a lexicalised reordering model. Moreover, a word penalty was introduced to the model for each generated English word, in addition to the language model; this factor is meant to bias towards longer output.

2.1.2. Lexical phrase-based translation with Moses

The process of training Moses on surface words only corpora is based on the same method as used by Pharaoh. Simply, no additional factors are trained on. For the purpose of this research, the baseline to factored models was trained with Moses in order to make sure we were comparing models trained in identical conditions. In particular, Moses did not support lexicalized reordering models at the time of this research, and so the baseline for factored models was did not use it.

2.2. Baselines error analysis

The evaluation stage was crucial to our understanding of how well the baseline models performed, and in what ways they could be improved. Evaluation was thus performed on either a sample or on the whole set of 2000 sentences translated by the various models. As we shall see, automatic methods are very useful for evaluating the output of machine translation and language generation systems in general. However, they do not replace the human editor's

sharp eye on what is an acceptable translation of the input and what is not. We thus used three different evaluation methods, the first of which was manual, and the two others automatic.

Manual evaluation was performed on a sample of the test set to give a precise account of the errors generated by the decoder, and in particular grammatical mistakes, which the automatic methods tend to miss. The BLEU metric, a standard metric for automatic output evaluation, was used to give an overall score on the translation quality. More fine-grained automatic evaluation methods were then implemented to complement the findings of the other two methods.

2.2.1. Manual output evaluation

Since our objective in this project was to improve the performance of existing translation models, we needed to get a detailed understanding of the types of errors generated by the decoder. Manual evaluation is tedious, time-consuming, and it is not always consistent: different human evaluators may disagree on the categorization of an error, or even on the relative "ungrammaticality" of a sentence or phrase. However, given the limitations of an automatic evaluation method such as BLEU (more on this in the section on automatic methods), manual evaluation seemed unavoidable to draw out the major issues that needed to be addressed in order to improve our models.

We shall first define the guidelines for manually evaluating a translation. Mainly, two types of translation errors were being tracked: ungrammaticality (or a lack of fluency) and failure in conveying the source text message (or a lack of adequacy and fidelity). It shall be noted at this point that the evaluation did not expect a one-to-one perfect translation of words and their inflection, although it was often the case that one word in the phrase was subject to mistranslation. Clearly, different languages often bear different ways of expressing the same idea, as it may be the case that, for instance, an original future tense is translated into the present tense for stylistic purposes. Therefore, a non-literal translation of a word or phrase's linguistic features was not necessarily penalized – and as we shall see, literal translations were actually often the source of mistakes – as long as the target sentence respected the standards defined above. We shall now explain what these standards mean. An ungrammatical sentence is one that does not conform to a language's grammar rules, and so it is quite easy to capture as the guidelines are clear: we are looking for gender/person/number agreement and overall syntactic coherence (i.e. the adjective comes before the noun in English; a sentence is composed of at least a verb and its irreducible arguments, etc.). A semantically correct

sentence is of course one that bears meaning, and in the case of a translated sentence, one that bears the same meaning as the original sentence did.

We decided to perform manual evaluation on the baselines trained on a limited portion of the Europarl corpus (50,000 sentences out of around 700,000) for French to English and English to French translations, for we were particularly interested in having a detailed report of the effect of our models trained on limited data. We thus extracted a sample of 150 sentences from those baselines and engaged in the endeavour of evaluating, counting and classifying ungrammatical and wrong translations. It should be noted that this work was performed by an experienced technical translator, thus alert to awkwardness in both languages.

Throughout this evaluation, several well-formed phrases or sentences that conveyed clearly the original message were encountered, which gave us an idea of the certain value of the current models. The focus of the evaluation, however, was on faulty translations, and on their importance in relation to the whole translation. The types of errors generated by the decoder had similar explanations in both French and English translations, but some were language-specific. Ungrammatical and/or semantically awkward cases in the output were determined, and categorized according to the assumed faults of the translation model in use:

- Source language words inserted in translation
- Wrong translations that rendered the translation not understandable without looking back at the source text
- Translations which had some relation with the original source text, but were considered wrong in different aspects
- Missing words (not translated from source text or missing in target text)
- Wrong reordering of syntactic components

The number of occurrences in each category was counted: this was meant to provide us with a metric as to how much editing needs to be performed by a human corrector on translation output before it could conceivably be released to the public. Possible shortcomings of the current model were then inferred in relation to these errors and to the specific target languages of the translation processes. Finally, some hypotheses have been designed as to how linguistically-enriched models could be used to decrease the error rate in each one of these categories.

2.2.1.1. Manual evaluation error report for Pharaoh lexical phrase-based English-to-French and French-to-English baseline models

Source language words introduced in translation

Pharaoh's current translation models do not include back-off or smoothing methods to deal with unseen words in the training data. Instead, Pharaoh simply inserts an unseen word as is into the translated output. It shall be noted too that the insertion of words in the source language into the output generally caused a bad syntactic order as well. For instance, in French to English translation, “il est temps de rompre avec une interprétation fallacieuse du principe de subsidiarité” (which could be translated as “ it is time we break up with a deceitful interpretation of the principle of subsidiarity”) was translated “it is time for an interpretation fallacieuse breaking with the principle of subsidiarity”, splitting the object of the sentence into two and introducing the verb in between. The problem of source language words inserted in the output translation mostly attests of a data sparseness problem; however, it requires some special attention because of the noticeable disruption it causes in the understandability of the translated text, and the frequency of its occurrence in our current models. These frequencies for French-to-English and English-to-French translations are shown in table 1.

Translation not understandable without looking at the source text

Phrases of this category had no relation whatsoever with the original text, thus forcing the reader/human corrector to refer to the latter. Such occurrences, like translating French verb “rassemble” (which could be translated as “brings together”) into “visiting” in English, were most likely the result of a wrong word alignment in the translation table, or, in some cases, they may be viewed as suitable translations in another context, but not in the one in question. See table 1 for the frequency of this error in Pharaoh baselines.

Wrong translations which had some relation with the original source text

These phrases were close to the original text, but either some translated words did not fit in to convey the original meaning (e.g. a wrong preposition like in “from greece to portugal” translated “de la grèce pour le portugal” instead of “de la grèce au portugal”), or the literal translation of the source text came up ungrammatical or meaningless in the target language (as in the wrong translation of “du bon sens paysan” – “the farmer’s good sense” – into “the right direction farmer”). Idiomatic structures often do not allow a one-to-one translation from source to target language, and if they have not been encountered in the training data (for the translation and the language model), it may be really difficult to get them right. Other syntax-

related issues which resulted in bad translations were word class ambiguity and a lack of agreement at the sentence or phrase level. This type of translation error is represented with its counts for both translation directions in table 2.

Literal translations

Syntactic structures such as negation, embedded clause structures, impersonal structures or reflexive structures are hard to translate because they often do not correspond in different languages, and translated syntactic arguments may need to undergo reordering (which we shall discuss in a later section), or a more or less important change ranging from word insertion or deletion to a complete transformation of the expression. For instance, the optional use of the pronoun “that” to open a relative clause, as in “an argument we can use”, is non-existent in French, where a pronoun is absolutely necessary; therefore the literal translation “un argument nous pouvons utiliser” is not acceptable. Likewise, a common noun in French always comes with a determiner, while in English this is not necessarily the case, as in the phrase “direct foreign investments”, which, when translated into French, should bear the indefinite determiner “des”. Another example of literal translation is the impersonal French expression “il faut” (“we shall”) into “it must” in English. Finally, the French negative form “ne ... pas” is tricky in that the verb is inserted in between the words conveying the negation, and has no corresponding structure with the English expression of the negation.

One advantage of phrase-based SMT (as opposed to word-based SMT) is the handling of fixed idiomatic expressions, when these occurred in the training data. Wrongly translated expressions were thus, for instance, English verbs with a fixed particle like “turned out”, “make up for”; these often suffered from a meaningless literal translation into French. In certain cases, the actual opposite of the original meaning was achieved, as in the translation of “le parlement ne présente plus” (“the parliament does not present anymore”) into “parliament presents more”. Finally, literal translation of fixed expressions often led to ungrammaticality in the target language, as was the case for the translation of “in favour of incorporating” into “en faveur d'inclure” (that can be transliterated as “in favour to include”); indeed, this expression expects a noun rather than a verb to follow.

As can be seen in table 2, this kind of error was very significant in both translation directions, and more so in English to French translation, where about 46% of such expressions were badly translated.

Same POS but wrong word

Expressions that include prepositions often do not correspond in French and in English. This issue is well-known to human translators, and so it is not surprising that automatic translation comes up against it. The choice of the wrong preposition by the decoder often caused ungrammatical or meaningless output. For example, in French to English translation, “elle est à l ' image de” was translated “it is for the image of”: both prepositions here are out of context and thus render the translation incomprehensible. In fact, “à l'image de” is just a prepositional phrase which means “like”. Similarly, “pour que l'élargissement” was translated “to that enlargement”, instead of “so that”. For English to French translation, phrasal verbs (verbs which take on a different meaning with a particle) often caused problem: “make up for lost time” was translated “qu'il fallait faire jusqu'à temps perdu” (transliterated “that needed to be done until lost time”), “to get my point across” was translated” obtenir mon point de” (transliterated “to obtain my point of”).

Wrong POS

Ambiguity at the lexical level led to ungrammatical and/or meaningless translated sentences or phrases. The phrase “ peuvent conseiller la bulgarie” (which could be translated as “can advise Bulgaria”) was translated by the decoder as “can bulgaria adviser”; here, the verb was mistakenly translated as a noun, thus turning the sentence ungrammatical. Wrong translation also led to wrong meaning. For instance, the auxiliary “have to” was sometimes translated into the verb “have”: “les agriculteurs ont tout simplement plus de savoir” takes on the mistaken meaning that the farmers have more knowledge, whilst the original meaning of that sentence was that the farmers “have to (or “must”) know”. It shall be noted at this point that part-of-speech translation can itself turn out to be ambiguous: in “I wish the negotiators continued success” translated as “je voudrais les négociateurs poursuivi avec succès” (transliterated “I wish the negotiators followed with success”), “continued” could equally be viewed in this context in the source language as an adjective or as a verb in the past participle. However, it should be translated into the target language solely as an adjective.

Wrong inflection

Often, the decoder translated successfully some words' lemma, but with the wrong inflection, like in “ j ' ai pu m ' en rendre compte” (which could be translated as “I was able to realize”) translated by the decoder as “i am able to realize”. In English-to-French translation, wrong inflections were much more widespread than in French-to-English translations (48 occurrences against 15), which could be expected: French being a morphologically richer

language than English, an English word may match more possible inflected words than in the other way round.

Wrong tense translations occurred for ambiguous tense forms: for instance, a verb base form in English takes on the same inflection in its present 1st or second person plural form, while this is not the case in French, giving rise to translations such as “les rapports [..] continuellement commenter” (transliterated “the reports continually to comment”) for the source text “reports [...] continually comment”, where the infinitive was attributed to the verb instead of its present tense form. Also, as mentioned earlier in the section on fixed syntactic structures, tense can be influenced by structural factors such as the verb's occurrence within a subordinate clause: the use of French subjunctive after the conjunction “que” was not respected twice, using the present tense instead.

Several occurrences of wrong agreements were found in both translation directions, mostly for English-to-French translations. Noun-modifier agreement (determiners, adjectives) were complicated by the non-existence of gender agreement in English, giving way to translations such as “la véritable poids” (where the determiner and the noun do not agree in gender) and “un aperçu très précises” (where the noun is masculine singular and the adjective is feminine plural). Mistakes in subject-verb agreement (such as in “i do not plays”) and in French translation subject-participle agreement (e.g. “mon rapport est également liée”, where the subject “rapport” is masculine and the present participle “liée” is feminine) often occurred in sentences which included embedded clauses. Finally, co-referencing pronouns translation such as “it” into masculine or feminine pronouns in French (and the other way round) caused problem: “the charter of fundamental rights because it summarises” was translated “la charte des droits fondamentaux parce qu'il summarises”, where the pronoun referencing to the first noun phrase should be feminine.

Missing words

Missing words in the output are words that appeared in the original text but were not translated into the target language, although they were essential to conveying the meaning of the original text. Missing words may be function words, but they can also be nouns and verbs. This phenomenon may have to do with wrong word alignment, or the most probable hypothesis selected for other reasons by the decoder did not include that word, or the translation selected may suit to a different context than the one in question.

Wrong order of syntactic components

For both translation directions, wrong ordering of syntactic components often occurred in long noun compounds, complex coordinated or embedded structures, but also on the sentence level in long, complex sentences, sometimes coordinated or separated by a comma. In some occurrences, the subject of the sentence was even interchanged with the object, as in “the fundamental rights which the public are entitled to” translated as “les droits fondamentaux qui ont droit à l'opinion publique” (transliterated “the fundamental rights which are entitled to public opinion”).

Distribution of errors – Pharaoh Baseline Error Report

	French-English		**English-French**	
	Total occurrences	Percentage out of total words in output	Total occurrences	Percentage out of total words in output
Source words inserted in translation	83	2.13%	79	1.77%
Wrong words translation	30	0.77%	17	0.38%

Table 1 : Translation Errors due to word alignment and sparse data

	French-English		English-French	
	Occurrences	Percentage out of total occurrences of same expression type	Occurrences	Percentage out of total occurrences of same expression type
Literal translations of fixed expressions (out of total fixed expressions)	31	14.8 %	52	46.84%
Same POS but wrong word (of total prepositions)	14	2.09%	13	2.64%
Wrong POS (of total nb of words)	11	0.28%	23	0.51%
Wrong tense (of verb phrases)	6	1.79%	14	4.37%
Wrong agreements (of verb agreement)	9	2.69%	Total: 34 Noun-modifier: 8 Subject-verb / participle: 23	2.76% (of total NPs and VPs) 0.87% (of NPs) 7.18% (of verb agreement)
Missing words (out of total words in output)	4	0.1%	9	0.2%
Total	**75**		**145**	

Table 2 : Translation Errors According to Various Linguistic Criteria[1]

[1] Percentages are computed for each case separately.

	French-English		English-French	
	Occurrences	Percentage out of total occurrences of same expression type	Occurrences	Percentage out of total occurrences of same expression type
Noun Phrases including adjectives (of all such NPs)	11	6.07%	17	6.31%
Wrong ordering at the sentence level (of total nb of clauses)	15	4.49%	10	3.1%
Total	**26**		**27**	

Table 3 : Wrong ordering[1]

Notes on error distribution computation

- Source word inserted in translation, wrong translations, words not translated and wrong POS errors were normalized by the total number of words in the output text.
- Fixed expressions that have been counted as correctly translated are the ones that do not have a literal translation from source to target language and nevertheless were correctly translated. This includes negation, impersonal structures, idioms, relative clauses that trace back to an external object and reflexive pronouns; these were manually estimated. In the French source text, there appeared to be around 209 cases of this type. For English, around 111. But these estimates were overall quite arbitrary, and it may be useful in future to redefine clearer guidelines for this type of error.
- Wrong preposition choice (same POS but wrong word) errors were normalized by the total number of prepositions in the source text.
- Wrong tense errors were normalized by the estimated number of verb phrases in the source text[2].

- Wrong agreements errors were normalized by the estimated number of verb phrases in the source text for translation into English; for translation into French, they were normalized by the total number of common nouns or by the number of verbs phrases in the source text.
- Correct coordinated noun phrases reordering was manually estimated.
- Wrong reordering on the sentence level was normalized by the number of clauses (including main and embedded clauses) in each sentence: this was estimated with the number of verb phrases in the source text.

2.2.1.2. Manual evaluation error report for Moses lexical phrase-based English-to-French and French-to-English baseline models

The distribution of errors for Moses French-to-English baselines is presented below. Because this model is very similar to Pharaoh baseline, we will not explain the details of this evaluation.

	Total occurrences	Percentage out of total words in output
French words inserted in translation	**79**	**2.03%**
Wrong words translation	**33**	**0.85%**

Table 4 : French-English: Translation Errors due to word alignment and sparse data

[2] In the English source text, verbs conjugated to the present and past tense, as well as modals were retained to represent verb phrases. Past participle, the gerund and the infinitival forms were thus discarded from this count. In the French source text, past and present participles and the infinitival form were discarded form the count. This was to avoid adding several times a phrase such as “I could have been liked” (where “have” is tagged as “VB” by the Brill tagger) to the count.

	Occurrences	Percentage out of total occurrences of same expression type
Literal translations of fixed expressions	34	16.26% (of total fixed expressions)
Same POS but wrong word	11	1.64% (of total prepositions)
Wrong POS	13	0.33% (of total nb of words)
Wrong tense	10	2.99% (of verb phrases)
Wrong agreements	6	1.79% (of verb agreement)
Missing words	33	0.85% (of total nb of words)
Total	**107**	

Table 5 : French-English: Translation Errors According to Various Linguistic Criteria[3]

	Occurrences	
Conjuncted Noun Phrases including adjectives	21	11.6%
Wrong ordering on the sentence level	14	4.19% (of total nb of clauses)
Total	**35**	

Table 6 : French-English: Wrong ordering[3]

2.2.2. BLEU evaluation

The advantages of using a standard automatic evaluation method for machine translation output are numerous. For one thing, it gives a relative comparison between different machine translation methods. As important is the ability to rate fast and at no cost the improvement of translation models. The BLEU metric automatically measures n-gram overlap with reference translations (the gold standard, or reference translation, is the human translation of the input

[3] Percentages are computed for each case separately

text in the target language). This method thus provides some representation for word choice and order. According to its authors, BLEU's unigram detection “tends to satisfy adequacy [while] the longer n-gram matches account for fluency” (Papineni, 2001).

Since it has been shown that the BLEU score correlates closely with human judgment, an improvement in BLEU score is commonly accepted as evidence for improvement in translation quality (Koehn, 2004). However, because it relies on n-gram comparison, the BLEU score gives credit to local word sequences rather than to a global translation quality. It has indeed been criticized for not allowing lexical variety (as different translations of a same text using different terms may be regarded as correct) and not accounting very well for grammaticality (Koehn and Monz, 2006). Overall, BLEU remains a widely-used metric for evaluating machine translation output.

2.2.2.1. Baselines BLEU scores

Trained on Corpus	Limited	Whole
French-to-English Translations	27.38	30.26
English-to-French Translations	28.08	32.42

Table 7: Pharaoh Lexical Phrase-Based Model

Trained on Corpus	Limited	Whole
French-to-English Translations	26.00	29.71
English-to-French Translations	21.86	27.73

Table 8: Moses Lexical Phrase-Based Model

The BLEU score was computed for our baselines by running the BLEU evaluation script on the decoder output and on reference translations of the target languages.

As was mentioned earlier, the baselines for Pharaoh and for Moses should not be compared to one another, as they were trained on different models. Moreover, we wish to emphasize the relative nature of the BLEU score: indeed, rather than being indicative in itself, it becomes interesting when compared to other models trained in the same conditions. Therefore, for instance, the fact that English to French translation for Pharaoh’s baselines got better BLEU scores than French to English translation, does not necessarily mean that English-to-French translation is an easier task. Because the BLEU score computes ngram matches, it is very sensitive among other things to tokenization, which may influence results. Other factors making the BLEU score a not entirely reliable metric are described in (Callison-

Burch, 2006). We should thus be very careful, when considering these results, to compare scores that are fairly comparable.

2.2.3. Fine-grained automatic evaluation

We designed alternative evaluation methods which could enable us to have a closer look at the sort of mistakes performed by the decoders, and their impact over the translation quality. Our evaluation method was designed to include various aspects of the translation which had proved to be significant problems in our manual evaluation, and was thus meant to expand on the latter, as well as to give a deeper understanding of the BLEU score's meaning. It involved the comparison of the source text to the translated text on the one hand, and of the translated text to the gold standard, on the other. The test set used was constituted of about 60000 words.

2.2.3.1. Source language words in output translation statistics

Foreign words inserted in translation were found to be one of the major problems encountered in both translation directions. We thus designed a method for automatically extracting and counting words in the output translation, which were identical to words of the same sentence in the original text. Our algorithm thus scanned simultaneously the decoder's output and the source text which had been used for translation, and checked if they bore identical words, which were then classified as "not translated".

Given the numerous expressions English has taken from Latin over time, some words which clearly had the same orthography in French and in English (such as words with –tion or –ure endings, like "action" or "candidature", without any French accents) and their plurals, were discarded from this list. We also discarded punctuation and numbers. However, several words such as proper nouns, which often bear the same orthography in both languages, were not discarded. The generated count of foreign words is thus more inclined to be informative when compared from one translation direction to another, rather than informative in itself. For a precise idea of that number, the reader may refer to the manual evaluation which was performed.

The source text compared with the output to be evaluated had been previously lemmatized; this information was kept together with the not translated words. By lemmatizing as well the French-English training corpus, we could thus compare lemmas of not translated words to those found in each translation model's respective source language training corpus. By counting words in the training data which had the same lemma as those not translated in the translation output, but with a different inflectional form, we intended to evaluate the

potential interest in using morphological information in the training data in order to counter the sparse data problem, which is assumed to be the main source to this "foreign word" phenomenon. Lemmatization, as opposed to stemming, provides the base form of an inflected word with the same part-of-speech, or syntactic function. Stemming provides the word's root, which includes other parts-of-speech from which that word may have been derived. For this primary research, we settled for more limited information and used Carroll's Morpha morphological analyzer and lemmatizer for English (Minnen et al, 2001), and Schmid's TreeTagger part-of-speech tagger and lemmatizer for French (1994). These tools shall be explained more in detail in the data linguistic preparation section.

The tables below present the statistics for source language words inserted in the baselines' output translations, and the number of word types that had morphological variants in the training data.

Source Language Words in Pharaoh Baselines

	French words (in percentage of tot. output words)	French words types	Word types that had a morphological variant in the training data
Limited training corpus (50,000 sentences)	6.15%	1923	1157
Whole training corpus (around 700,000 sentences)	3.98%	851	442

Table 9 : French to English Translations

	English (in percentage of tot. output words)	English words types	Word types that had a morphological variant in the training data
Limited training corpus (50,000 sentences)	4.73%	1656	1082
Whole training corpus (around 700,000 sentences)	3.29%	806	682

Table 10 : English to French Translations

Results are similar in both translation directions, which seems reasonable since both models were trained on the same parallel corpus, and Pharaoh's decoding was performed on human translations of a same text.

On a limited training corpus size, it seemed harder to translate from French; about 6% of the whole translation into English was composed of not translated French words, while in English to French, such was the case for around 4.7% of words. The morphological diversity of French words appeared to be an important cause to this: around 60% of word types not translated had a morphological variant in the training corpus. For English to French translation, this ratio was higher still, but on a more limited number of word types.

This tendency to a more difficult translation from French was blurred when increasing the training corpus size. Increasing the training corpus by 10 greatly helped reduce the number of source words inserted as is in the target output.

Source language words in Moses baselines

	French words (in percentage of tot. output words)	French words types	Word types that had a morphological variant in the training data
Limited training corpus (50,000 sentences)	6.08%	1905	1142
Whole training corpus (around 700,000 sentences)	3.9%	847	438

Table 11 : French to English Translations

	English (in percentage of tot. output words)	English words types	Word types that had a morphological variant in the training data
Limited training corpus (50,000 sentences)	4.88%	1667	1094
Whole training corpus (around 700,000 sentences)	3.39%	690	818

Table 12 : English to French Translations

The results on foreign words inserted in the output translation show that Moses baselines for both translation directions perform similarly to Pharaoh baselines, as expected because both models are based on the same principle. The lack of use of lexical reordering for Moses models may have influenced the actual ordering of words (thus explaining partly the lower BLEU scores that these models obtained), but it seems that Moses performs about as well as Pharaoh when trained on surface words only.

2.2.3.2. Mistranslated words categorization

To compare automatically-generated text to a gold standard, several techniques may be used. (Popovic et al., 2006), for instance Word Error Rate, a widespread technique for evaluating speech recognition performance, which computes word insertion, deletion and substitution. The BLEU method, as described earlier, computes ngram matches between the output and reference texts. We decided to focus on word deletion and expand on the most frequently mistranslated word classes compared to our gold standard: a word that appeared in the gold standard but did not in the exact same form in the decoder's output was considered mistranslated. A word that appeared in the gold standard and also appeared in Pharaoh's output, but with a different inflection, was categorized as mistranslated, but with the right lemma. We thus compared the decoder's output to the gold standard in the corresponding language, sentence by sentence. Both texts were POS-tagged using the Brill tagger (Brill, 1995) for English, and the TreeTagger for French, and lemmatized using the same tools mentioned above.

The comparison was performed starting from the reference sentence, and all percentages were normalized by word numbers from the reference text. This way, we could find out the distribution of part-of-speech that had been mistranslated. The word comparison was position-independent, and to make sure a word in the output text was not used more than once, it was removed from the bag of words left for each sentence, once recognized in the corresponding reference sentence. Reference words left unrecognized were then scanned for their lemmas against the output lemmas left. Finally, those words which were not recognized, either for their word form or their lemma, were added to the list of "mistranslated words". All results were outputted categorized by part-of-speech, with their number of occurrences and their ratio in relation to the total number of words of that POS category in the reference text.

Mistranslated Words in Pharaoh Baselines

All results in the following tables are percentages out of the total number of words for that same POS in the reference text (unless stated otherwise).

French-to-English translations

	Conj CC	Num CD	Det DT, PDT, WDT	Prep IN	Adj JJ, JJS, JJR	Mod MD	Noun NN, NNS	Pron. PRP, PRP$	Adv RB, RBR, RBS, RP	To TO	Verb infinitive VB	Verb inflected VBD, VBG, VBN, VBP, VBZ	Wh pro WP, WP$, WRB
Mistranslated words by POS	17.7	32.8	24.8	40.2	46	**54.7**	38.5	34.7	**54**	37.8	**60.2**	**71.8**	49.1
Right lemma but wrong inflection	-	-	-	-	0.5	-	**3.3**	-	-	-	**8.3**	**17.6**	-
Total not translated	**38.9 % (**out of total words in output translation)												
Total right lemma	**2.65%** (out of total words in output translation)												

Table 13 : French to English Translations - Limited training corpus (50,000 sentences)

	Conj CC	Num CD	Det DT, PDT, WDT	Prep IN	Adj JJ, JJS, JJR	Mod MD	Noun NN, NNS	Pron PRP, PRP$	Adv RB, RBR, RBS, RP	To TO	Verb infinit VB	Verb inflected VBD, VBG, VBN, VBP, VBZ	Wh pro WP, WP$, WRB
Mistranslated words by POS	18.3	33.3	24.6	**38.6**	**41.8**	**52.5**	34.6	33.3	**51.4**	35.2	**54.3**	**66**	**46.5**
Right lemma but wrong inflection	-	-	-	-	0.4	-	**3.4**	-	-	-	**8.7**	**18**	-
Total not translated	**36.1% (**out of total words in output translation)												
Total right lemma	**2.7%** (out of total words in output translation)												

Table 14 : French to English Translations – Whole training corpus (~700,000 sentences)

As expected, the ratio of mistranslated words in Pharaoh's output for French-to-English translation clearly decreased when the corpus size was increased. The ratio of mistranslated words that had the right lemma represented about 2.7% of all translated words, which is quite high.

In the limited trained model, words that were most often mistranslated were verbs, with a high portion of wrongly translated inflected verbs; then, many function words were mistranslated: adverbs, modals (again reflecting the idea that the tense may have been wrong in the output sentence), WH- pronouns and prepositions; finally, adjectives and nouns. This order was about the same in the whole corpus trained model. It thus seems that the problem we had noted in our manual evaluation regarding prepositions also applied to other function words. However, (and this is one of the disadvantages of this automatic techniques), it is hard to say whether the mistranslated words of each category were translated into a correct synonym or not. What we can say is that a rather small portion of them was translated, but in the wrong inflectional form: 18% of inflected verbs in the model trained on the whole corpus

had the wrong inflection, 8.7% of infinitive verbs and 3.4% of nouns. We had a closer look at the actual morphological variants that our models proposed instead of these: singular nouns were mostly output in their plural form and vice-versa, sometimes as verbs (e.g “hope” was translated “hoping”). Infinitival verbs were output in any of their inflected forms, while inflected verbs were output either as in their infinitival form, or in another inflected form, sometimes as nouns.

English-to-French translations

	Numbers NUM	Determiners DET	Prepositions PRP	Adjectives ADJ	Nouns NOM	Pronouns PRO	Adverbs ADV	Verbs VER
Mistranslated words by POS	26.5	15	39	39.5	34.8	37.9	50.24	53.3
Right lemma but wrong inflection	0.5	36.8	7.7	11.6	2.7	16.2	2.72	16.4
Total not translated	**37.2% (**out of total words in output translation)							
Total right lemma	**11.6%** (out of total words in output translation)							

Table 15 : English to French Translations - Limited training corpus (50,000 sentences)

	Numbers NUM	Determiners DET	Prepositions PRP	Adjectives ADJ	Nouns NOM	Pronouns PRO	Adverbs ADV	Verbs VER
Mistranslated words by POS	25	14.4	35.8	34.9	30.6	36.2	48.6	49.2
Right lemma but wrong inflection	-	35.8	7.6	11	2.5	16.1	2.16	16
Total not translated	**34.1% (**out of total words in output translation)							
Total right lemma	**11.3%** (out of total words in output translation)							

Table 16 : English to French Translations – While training corpus (~700,000 sentences)

In English to French translation, a large part of mistranslated words had their lemma rightly translated, but with the wrong inflection: such was the case for 11.3% of translations for the model trained on the whole corpus.

In the limited trained model, words that were most often mistranslated were verbs, adverbs, pronouns (which can be either function words as “qui”, or personal pronouns), prepositions and finally adjectives. This order of importance in the errors generated is similar to the one found in French to English translations. The distribution of right lemmas with wrong inflection are however different: wrong determiner inflection with the right lemma represented 36.8% to 35.8% of determiner occurrences (for models trained on limited and whole corpus respectively). Determiners are much more widespread in French than in English, and also, if the existing set of determiners is rather limited, it involves all possible inflections (singular, plural, feminine, masculine). Mistranslated adjectives were usually translated into plural form when they should have been singular, into feminine form when they should have been masculine, and vice versa. Verbs were translated in various inflected forms, given the many conjugation possibilities in French.

Although these findings do not clearly establish that these translations were wrong from a grammatical or semantic point of view, they provide some information on the types of mistakes performed, and support the idea that using linguistic knowledge in our models may help resolving an important part of the mistranslation. They also highlight the difficulty of

generating the right inflection for a word that did not exist in the input of the translation process.

Mistranslated Words in Moses Baselines

All results in the following tables are percentages out of the total number of words for that same POS in the reference text (unless stated otherwise).

French-to-English translations

	Conj CC	Num CD	Det DT, PDT, WDT	Prep IN	Adj JJ, JJS, JJR	Mod MD	Noun NN, NNS	Pron PRP, PRP$	Adv RB, RBR, RBS, RP	To TO	Verb infinitive VB	Verb inflected VBD, VBG, VBN, VBP, VBZ	Wh pro WP, WP$, WRB
Mistranslated words by POS	17.8	35.7	24.1	**41.2**	46.3	56.4	**39.4**	35.9	55.3	36.9	**60**	**72.2**	51.4
Right lemma but wrong inflection	-	-	-	-	0.3	-	**3.5**	-	-	-	**8.6**	**18**	-
Total not translated	39.3**% (**out of total words in output translation)												
Total right lemma	2.7**%** (out of total words in output translation)												

Table 17 : French to English Translations - Limited training corpus (50,000 sentences)

	Conj CC	Num CD	Det DT, PDT, WDT	Prep IN	Adj JJ, JJS, JJR	Mod MD	Noun NN, NNS	Pron PRP, PRP$	Adv RB, RBR, RBS, RP	To TO	Verb infinit VB	Verb inflected VBD, VBG, VBN, VBP, VBZ	Wh pro WP, WP$, WRB
Mistranslated words by POS	17.9	31.3	24.9	**38.9**	42.3	51.2	**35**	33.2	51.3	34.3	**54.2**	**65.9**	48
Right lemma but wrong inflection	-	-	-	-	0.3	-	**3.5**	-	-	-	**9.1**	**18.54**	-
Total not translated	36.2**% (**out of total words in output translation)												
Total right lemma	2.7**%** (out of total words in output translation)												

Table 18 : French to English Translations – Whole training corpus (~700,000 sentences)

English-to-French translations

	Numbers NUM	Determiners DET	Prepositions PRP	Adjectives ADJ	Nouns NOM	Pronouns PRO	Adverbs ADV	Verbs VER
Mistranslated words by POS	23	16	**37.8**	38	**36.3**	36.8	38	**53.8**
Right lemma but wrong inflection	0.5	**29.8**	2.8	13.2	2.9	**13.3**	0.2	**17.3**
Total not translated	37.4**% (**out of total words in output translation)							
Total right lemma	9.6**%** (out of total words in output translation)							

Table 19 : English to French Translations - Limited training corpus (50,000 sentences)

	Numbers NUM	Determiners DET	Prepositions PRP	Adjectives ADJ	Nouns NOM	Pronouns PRO	Adverbs ADV	Verbs VER
Mistranslated words by POS	24	14.4	**37.4**	36.1	**30.7**	37.6	48.3	**51.5**
Right lemma but wrong inflection	-	**35.8**	4.5	13.2	2.7	**14.1**	1.2	**16.8**
Total not translated	35**% (**out of total words in output translation)							
Total right lemma	10.6**%** (out of total words in output translation)							

Table 20 : English to French Translations – While training corpus (~700,000 sentences)

2.2.4. State-of-the-art phrase-based SMT performance

Several problems were encountered in our baselines that affected translation quality. First of all, the BLEU scores clearly demonstrated that a larger corpus improved translation quality, which was also supported in our more fine-grained automatic method: the larger the corpus, the less source language words inserted in the output translation, and also the less mistranslated words as compared to a gold standard. Indeed, a small corpus created a data sparseness problem for our statistical models, including unseen words in the training data and bad word alignment in the translation model (the parallel corpus on which the model was trained involves human translations of a same corpus, which can be expressed in very different ways and thus complicate the alignment, especially if there is a shortage in training data). A smaller corpus also leads to unseen ngrams in the language model.

Together, these problems were responsible for the insertion of not translated words as is from the source language into the target language text – which was found to be a major issue in our manual evaluation – and for the selection by the decoder of unfit translation hypotheses for certain words and word sequences. Indeed, unseen word sequences in the source language text cause the decoder to translate smaller available phrases, which do not necessarily take on the right meaning (if any at all) when joined together. This was particularly problematic, as we have seen in our manual evaluation, for structures involving

fixed particles and for compositional phrases, even more so if these did not hold a literal correspondence to their translated counterpart: the literal translation problem was found to be one of the most common in both translation directions. The issue was complicated by the presence of embedded clauses and modifiers within rigid structures, which are unlikely to have been seen in the training corpus. The same principle worked for more complex syntactic structures such as coordination, which often caused wrong ordering of syntactic components.

Data sparseness was not only a consequence of the training corpus size, but also of the model simplification: because the baseline models were trained on surface words (i.e. lexical items) only, data sparseness was more likely to happen than if the models had been trained on a more general representation of the word (e.g. lemma or part-of-speech). Word inflection and part-of-speech ambiguity were found to be quite problematic in both translation directions, but much more so when translating into French, for the reasons we have seen. We have argued that integrating linguistic information to our current models may improve grammaticality in both French and English outputs: for instance, it was often the case that the models translated the right lemmas with the wrong inflectional form, or that a word was mistranslated due to part-of-speech ambiguity. In the next sections, we propose some models which take these issues into account.

Chapter III: Methodology

In accordance with the findings of the previous section, possible features were envisaged to enrich the current model, and experiments were designed for the two suggested approaches.

3.1. Experimentation design

As we have seen, most of the translation errors mentioned in the previous section have at least one common solution, and that is to have more training data. As we know, for statistically-trained models, “there is no data like more data”, and so by adding linguistic occurrences to our pool of knowledge, we may augment the probability of getting the right translation given an input. However, getting parallel corpora is not straightforward, and for some languages limited data is readily available.

In this thesis, we wished to focus on how linguistically-informed models may contribute to getting better translations for English-to-French and French-to-English translations from several standpoints: a better general readability and understandability of a document should be achieved mainly by 1. Using exact terminology, 2. Assuring a correct ordering of syntactic components and 3. Assuring correct inflection of the translated words in context. We hoped to achieve these improvements by enriching the data at hand, especially if this data was scarce. Errors discovered throughout the evaluation stage were thus analyzed and research hypotheses formulated with respect to possible linguistic features that may aid the resolution of such errors. The hypotheses cover both translation directions, as the problems encountered were often assumed to have similar origins, but each language's specificity was taken into account.

While building hypotheses, we also had to consider how this linguistic knowledge should be integrated in the proposed models – combined feature models and factored models. These two models are similar from a theoretical standpoint, with two major differences: on the one hand, while translating features from source to target language was made possible by the combined feature models, thus making use of linguistic properties of the input, our factored models would rely solely upon surface word translation. On the other hand, to ensure grammaticality and coherence in the target language, both models relied mainly upon linguistically augmented language models, which for the factored models were relatively independently trained, while such was not the case for combined feature models. These two

factors had to be taken into account while creating the hypotheses. In general, linguistic feature modelling was based upon the same concepts, but implemented in different ways.

3.1.1. Source language words inserted in translation

The unseen word problem in Pharaoh and Moses output translations had an important role in decreasing the text readability, and was thus carefully studied in the framework of our evaluation. Given the models we wished to experiments, we did not, however, plan to find a solution to this problem. This may well involve the design of a backoff mechanism using more general linguistic features than the lexical level which Pharaoh is trained on, as was done by Kirchhoff and Yang (2006), but this is out of this work's scope and shall be discussed as a potential future integration to our models.

3.1.2. Word alignment

As was mentioned in the error report, it was assumed that such extreme translation mistakes where the output had no relation with the source text had to do in most cases with mistaken word alignment. As we have seen, improving word alignment with morphological normalization and morpho-syntactic information have been performed successfully in the past years. However, in the framework of this research, we did not focus on improving the word alignment.

Indeed, it was not possible to generalize the word alignment in our combined feature models to, say, lemma information because the lexical information of a word was a necessary component of our combined models, outlining one limit of this type of model. This way, we could keep track of it throughout the translation process, and finally recover this essential linguistic level in the decoder’s output. Generalization of our models could not be solved either by combining, say, only the lemma and morphological features, since generalization requires a representation of the word which covers several word forms. Finally, combining surface words and their lemmas may help disambiguating the few homographs which have the same part-of-speech, but different lemmas (and thus different meanings), but certainly not provide extra information to the decoder as to which words are of the same family, putting down the possibility of improving cases in which the translation had no relation whatsoever with the source input.

On the other hand, factored models potentially provided an interesting twist for improving translation: given the separate modelling of lemmas and surface words in factored models, and the fact that the same lemma may well have been encountered more often in the

training data than the surface words it represents, we hoped that the decoder would be biased towards choosing right translation for words of a same family. At this stage of the research, this was hypothesized to be a good compromise for not dealing with word alignment improvement.

3.1.3. Wrong translations which had some relation with the original source text

3.1.3.1. Literal translations

The problem with literal translations was that they often produced ungrammatical sentences in the target language. We hypothesized that the use of POS information in language models would help:

- Limiting the generation of ungrammatical sentences that involve the unwanted insertion of translated source language words or word sequences in the target output, or, on the contrary, the lack of translation of essential words. To illustrate, the probability of ending a French sentence with a preposition is much weaker than in English, and our new ngram models – either when enriched with POS information in the case of combined feature models, or when trained on this factor alone in the case of factored models – would include this information to some extent. It may also prevent sentences with two verbs as were encountered in our baselines. Fixed expressions like the negation “there can be no” (translated “il peut y avoir aucune”, missing out the negative “ne” particle) may be captured by a wider ngram than the current trigram language models in use.
- Influencing the correct output of expressions which require to be followed by a certain part-of-speech such as “en faveur de” (which expected a noun and not a verb, as it was the case in our English-to-French translation Pharaoh baseline). For combined models, it may be the case that this expression had been encountered in the training data followed by nouns; models used by the decoder may then bias it to choose those nouns that followed this expression in the training corpus, rather than translating the following word separately. As for factored models, they may learn, for instance, that a preposition, noun and preposition sequence in French is more likely to be followed by a noun.

It was thus envisaged to use the surface word combined with its part-of-speech in both translation directions, with a trigram language model, for the combined feature models. As for factored models, we planned to train a POS factor on top of our surface word model.

Fixed expressions which require a certain amount of transformation were harder to handle: getting from “ne doit dépendre que de” (transliterated “not must depend only on”) to “should depend only on” is a large step if this precise expression was not encountered in the training data. Also, the choice of the right word among a bag of candidates with the same type of syntactic function (in our evaluation, we presented the case of prepositions) showed tricky when each one of them could have been translated in different ways according to the idiomatic context. These issues were thus left aside.

3.1.3.2. Wrong POS

Clearly, it was expected that the addition of the POS feature/factor to the models would help disambiguate a word to get its appropriate translation in context: for instance in “has underlined this once more”, the pronoun “this” is followed by an adverb, and so it is unlikely to be translated (as this was the case in our evaluation sample) as the determiner “le”, which is most likely to precede nouns. A deeper syntactic analysis, however, may also be of use, for example in the case of “continued success” previously described, and for other more complex structures.

3.1.3.3. Wrong inflection

If we wished to assure the correct inflection of words in context in the target language, we needed to concentrate on several issues: ideally, the inflectional features should be translated from one language to another, and agreement in the target language should be assured in context and using language-specific features.

For combined feature models, it was hypothesized that translating inflectional features of words may be addressed by combining surface word, POS and morphological information: for wrong agreements within phrases such as “these recent bus”, a combined model representing the pronoun “these” in its plural form and the noun “buses” in its plural form may bias the decoder towards a correct agreement of the phrase head and its modifier. But the combination of morphology and surface word may suffice. Regarding agreement in longer phrases where, for instance, the verb is separated from the subject by embedded clauses (which resulted in our Pharaoh baseline in translating a present form into a base form from English to French), this may be solved using larger language models bearing the inflectional

information. Given the sparse data problem inherent in the combined feature model, we hoped that bigger n-grams made possible in the factored models (trained on morphology and POS separately) would help.

However, some issues should be treated in a wider context. For instance, structural constraints may influence a verb's tense, as it is often the case in the French subordinate clause opened by the conjunction “que” that the verb takes on the subjunctive form. We thus envisaged creating a chunk feature to replace words in their syntactic context. In fact, it seemed likely that using chunk and morphological information, both in combined and in factored models, may help ensure agreement, especially in cases where the agreed elements are not next to each other.

Finally, it is essential to note that, as French is a morphologically-richer language than English, it would definitely pose some problems of its own. The more numerous possible tenses in French, as well as morphological inflections for certain parts-of-speech which are non-existent in English (such as gender for determiners, adjectives, and in verbal inflection in the 3rd person), constitute a wider bag of potential translations from which the decoder can choose for English-to-French translation. Some agreements in French, such as subject-participle agreement, do not exist in English either. It was hoped that the morphologically-informed model would limit the corresponding tenses available for the translation of certain English tenses, and that, combined with (or in parallel to, if we are talking about factored models) deeper analysis features such as POS and chunk, this would allow correct agreements in French. The problem of co-reference outlined in the error report was not expected to be solved by these models.

3.1.4. Wrong order of syntactic components

The wrong ordering of syntactic components was mostly blamed on the reordering model failure. The use of POS information in our models could possibly improve word ordering, especially in the case of noun compounds and coordinated noun phrases including adjectives. Ordering on the sentence level could probably be improved by the use of the chunk feature/factor. The language model based on it may bias the models towards certain orderings of chunks, especially coordinated chunks which was a major problem.

3.1.5. Models for experimentation

The following models were designed for implementation in both translation directions for combined feature models:

1. To deal with ambiguity at the lexical level, literal translations and missing words:
    - Word + lemma
    - Word + POS
2. To deal with wrong inflections:
    - Word + morphology
    - Word + POS + morphology
    - Word + chunk

The following models were designed for implementation in both translation directions for factored models:

1. To deal with ambiguity at the lexical level, literal translations and missing words:
    - Word + lemma
    - Word + POS
    - Word + POS + lemma
2. To deal with wrong inflections:
    - Word + morphology
    - Word + POS + morphology
    - Word + chunk

Implementing the designed experiments involved enriching the data at hand with linguistic information. Various existing NLP tools were considered and evaluated to annotate the French and English corpora: features outlined in the previous section included POS tagging, word lemmatization and morphological analysis, chunking. In the next step, experiments were implemented, using the two different linguistically-enriched models proposed, to try and test the hypotheses.

3.2. Building linguistically-informed models for MT

3.2.1. Data preparation

The data preparation stage was very similar for both intended models: we needed to create a parallel corpus for French and English that would incorporate linguistic knowledge. To this aim, we suggested to linguistically process the Europarl corpus with off-the-shelf NLP tools, and to then adjoin their output to words in the corpora of the corresponding languages. The intended format was the following:

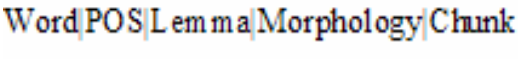

Figure 3 : Linguistically-Enriched Corpus Format

The word is the lexical representation of the word: this is the level that our baselines were trained on. POS information is the word's part-of-speech tag in the context of the sentence. The lemma is the word's base form of the same part-of-speech category (as opposed to its stem form). The morphology is the word's inflectional information, and the chunk its syntactic role in the sentence.

This information would be treated differently by our models: combined feature models would regard the above as one entity or token, while the generation models used in factored models would be able to process these linguistic levels separately. Therefore, we needed to generate separate corpora for each intended experiment with the combined feature models, by extracting the necessary features from the above main representation, while for the factored models, a single augmented parallel corpus could be used by simply defining the factors to train on. Additional corpus representations, each bearing one linguistic level, should be extracted to train on them language models.

3.2.1.1. Feature design

Various considerations were taken into account to choose the appropriate tools for natural language analysis, including the level of detail of their output and its relevance to the task envisaged, and the tools' level of performance on unseen data. Since tokenization and tagsets are different across languages, we needed tools for French and English corpora analysis that had been trained on each language respectively.

French features

POS tagging

Several freely available tools exist for POS tagging for French. We chose Schmid's TreeTagger, a probabilistic POS tagger adaptable to several languages which uses decision trees (Schmid, 1994). This method defies commonly used n-gram taggers and predicts tag sequences by building probabilistic decision trees: the authors claim that this method is more adapted for dealing with sparse data than Markov Model based taggers. The tagger came ready with parameter files for French and thus did not need to be trained, which was an

advantage as we did not possess a POS-tagged corpus. This also meant that the tool was trained on a different domain than the corpus we were intending to tag, which could decrease performance. However, the tool was reported by its developers to have reached accuracy rates of above 94% on unseen data, and to handle sparse data problem by using a limited built-in lexicon with a-priori tag probabilities and performing additional suffix analysis. The TreeTagger performs text tokenization and outputs one word per line with its part-of-speech and its lemma. The tagset used by TreeTagger for French was composed of 32 tags, which did not include morphological information such as gender or number:

ABR, ADJ, ADV, DET:ART, DET:POS, INT, KON, NOM, NUM, PRE(1rst), PRO, PRO(DEM), PRO(IND), PRO(PER), PRO(POS), PRO(REL), PRP, PRP(det), PUN, PUN(cit), VER(aux), VER(cond), VER(futu), VER(impf), VER(infi), VER(pper), VER(ppre), VER(pres), VER(simp), VER(subi), VER(subp), SENT.

Figure 4 : Tagset for French POS Tagging

Lemmatization and morphological analysis

As mentioned in the previous section, the TreeTagger for French already provided the lemma. Another tool for French text processing, Flemm (Fiametti, 2000), takes as input a word and its POS tag as provided by the TreeTagger, and outputs non-contextual morphological analysis as well as the word's lemma. Another of the tool's functionalities involve checking the POS tag against the word's suffix, and possibly corrects wrong tags.

The morphological tags are coded according to the lexical specifications recommended for French by the Multext consortium (Véronis, 1996).

Nouns	**Cat**	**Type**	**Gender**	**number**				
	N	c,p	m,f	s,p				
Verbs	**Cat**	**Type**	**Mood**	**tns**	**pers**	**nb**	**gender**	**group**
	V	m,a	i,s,m,n,p	p,i,f,s	123	s,p	m,f	123
Adjs	**Cat**	**Type**	**Gender**	**number**				
	A	f,o,i,s	m,f	s,p				
Pros	**Cat**	**Type**	**Person**	**gender**	**number**	**case**	**Poss**	
	P	p,d,i,s,t,r,x	123	m,f	s,p	n,j,o	s,p	
Dets	**cat**	**Type**	**Person**	**gender**	**number**	**poss**	**Quant**	
	D	a,d,i,s,t	123	m,f	s,p	s,p	d,i	
PrepDets	**cat**	**Type**	**Gender**	**number**	**Quant**			
	Sp+D	A	m,f	s,p	D			

Figure 5: Tagset for French Morphological Analysis

Please see the appendices for more information on the meaning of value codes. The morphological analysis of Flemm thus included the part-of-speech of the word; we had to

decide whether we'd keep it in our feature or discard it, and whether the morphological features should be kept together, or decomposed. We decided to keep the POS information in the morphological feature, given the very different combinations of values that may be assigned to different parts-of-speech, as we wished the language model to represent word agreement; this would not undermine out experimentation with the separate POS feature, which includes other tags for words that have no particular morphological analysis. For these primary experiments, we also decided to keep the morphological features as one block that would be the equivalent to its less complex English counterpart.

Flemm is a mainly rule-based system: it uses a few hundred rules and a list of exceptions, but no lexicon. It performs morphological analysis by segmenting the word and considering its extension according to its part-of-speech: for example, if a word is an inflected verb, possible endings to this verb are considered, and following, the possible segmentation of the word base and of its inflection are evaluated. The fact that Flemm does not use a lexicon means its analysis is relatively robust, but it occasionally produces wrong lemmas. Moreover, because it is non-contextual, it presents all the possible morphological inflections for a word, which for our purpose was problematic: we wished to limit as much as possible the possible translations of word sequences from source to target language in our translation models. Also, from a technical point of view, our features could only bear one possible tag. Flemm's output thus required post-processing.

Rule-based morphological disambiguation for French

The best solution to the morphological ambiguity problem would have been to create a statistical automatic morphological disambiguator based on word sequences. Unfortunately, we did not possess a morphologically tagged corpus to train on. We thus decided to have a closer look at the types of ambiguities generated by Flemm and to manually create rules choosing the most probable tags for ambiguous words. We noted two main ambiguity types: inflectional ambiguities and lemma ambiguities.

Inflectional ambiguities

Verbs (person, tense, mood)

Several verb inflections in French take on the same orthographic form, while they represent a different tense and/or person. For example, the inflected verb form "signifie" ("signifies") could be seen either as a first or third person singular in the present indicative, as a second

person singular in the imperative present, or as a first or third person singular in the subjunctive present.

Pronouns (type, case)

In French, the first and second person plural pronouns (“nous” and “vous”) are used as personal and reflexive pronouns, which creates occurrences such as “Nous nous interrogeons” (“we ask ourselves”). Another issue is with pronoun case, where for instance the third person feminine singular “elle” can be either subject or indirect object to a sentence, as in “Elle parle mal d’elle” (“She speaks badly of her”). The cases described above thus influence the pronouns’ syntactic role classification (the POS tagger does not differentiate the nuance), as well as their lemma: the lemma for “elle” could be either “il” or “lui”.

Lemma ambiguity (homographs)

Verbs

Some inflected verbs are ambiguous from the viewpoint of their lemma. For example, in the phrase “m'a vraiment plu”, “plu” could be seen as the verb “plaire” (“to please”) or the verb “pleuvoir” (“to rain”) in the past participle form.

Pronouns

The lemma ambiguity for pronouns was described above.

Nouns

Some nouns that are written the same way are ambiguous. For example, in the phrase “au cours de” (“in the course of”), the word “cours” could be viewed as the noun “course” in its singular form, or the noun “cour” (“playground”) in its plural form.

Hypotheses on morphological patterns

To create the rules that would automatically select one possible analysis out of the ones proposed for ambiguous cases, we first came up with a few hypotheses:

- Inflectional ambiguities are consistent among verbs (or auxiliaries) of a same group. In French, 3 main groups exist: verbs with –er ending (like “aimer”), verbs with –ir ending (like “finir”) and all the other verbs, called “irregular verbs”.
- Inflectional ambiguities can be similar from one group to another given a same mood, tense and person.

- The domain of the training corpus allows a few hypotheses regarding the tenses and persons which are more or less likely to occur. For instance, “tu” is not very likely to occur in a formal type of corpus.
- Some morphological distinctions in certain cases are less important for the specific task of French to English and English to French translation.

We then performed a manual analysis of a section of Flemm’s output on our French corpus and verified if our hypotheses were true. It appeared that the first hypothesis was justified: verbs of a same group behaved identically when inflected, even for verbs of the third group which involves several different extensions. However, hypothesis 2 proved to be wrong. We thus needed to keep a clear distinction between verbs of different groups, and to relate to auxiliaries as a separate group as well. Hypotheses 3 and 4 allowed us to reach some decisions regarding some morphological analyses which could be set aside, as they did not seem to influence the MT task for this specific corpus and the specific languages in question.

Rules for morphological disambiguation

Based on our findings, we thus proposed several rules: most of them constituted an acceptable compromise for limiting the ambiguities, others were much more arbitrary.

First of all, we documented the most widespread inflectional ambiguities for verbs, and where possible, we designed modified versions of the morphological tags that limited the level of ambiguity. For instance, for the verb “aller” (“to go”) of the third group, in its present first person plural form “allons”, could be either indicative or imperative: we thus included both possibilities in the tag, so that “Vmip1p” – standing for “verb, type: main, mood: indicative, tense: present, 1rst person, number: plural” – became “Vmi/mp1p” – where the notation “i/m” includes both indicative and imperative moods. This particular rule appeared to be mostly true for all three groups, not including auxiliaries which were thus discarded from it: most of the time in French (apart from very rare cases), the indicative present first person plural form of a verb is the same as its counterpart in the imperative mood. This rule, together with several others of the same type, was thus considered as relatively safe. Its gain in the bigger framework of our future translation models seemed also satisfying: a verb form could hopefully be translated into different corresponding English verb forms with regard to context.

In the context of the corpus domain in question, some fairly straightforward decisions could also be taken. For instance, because our corpus was mostly in direct speech, the inflected verb form “suis” was more likely to be the present first person singular form of the

verb “être” (“to be”) than of the verb “suivre” (“to follow”). Also, we completely discarded the second person singular present indicative and imperative moods, which were very unlikely to occur in a formal type of corpus.

Pronoun disambiguation was somehow more problematic, as it is really a context-dependent issue, and pronouns were tagged as personal pronouns, whether they were nominative or object, personal or reflexive. We thus assigned quite arbitrarily what we thought was the most frequent tag, i.e. for example, third person singular pronoun “elle” was indifferently assigned a nominative case, which limits the information, in the view of our translation models, of whether it should be translated as “she” or as “her”. The same principle was applied to pronouns’ lemmas, where for example the first and second person plural pronouns were assigned the lemma for personal pronouns “il” (when they could also be reflexive pronouns).

Finally, some problems were left unsolved; in these cases, the first analysis proposed by the morphological analyser was picked just for the purpose of assigning only one analysis per word. These cases included ambiguous nouns, words had been wrongly POS tagged such as homographs (like in “y compris”, which is an adverb, but was wrongly tagged as the past tense conjugated verb “je compris”) and typo mistakes (such as “euxmemes” where the two words should be separated by a dash). For a more detailed account of all the rules designed, please refer to the appendix.

Morphological disambiguation evaluation

At the bottom line, inflections should ideally be chosen according to context, and so an automatic method rather than a rule-based one seems more appropriate for this task. However, we performed manual evaluation on 150 modified tags in context, and it appeared that our method was quite reliable: 118 cases were correctly tagged in context. Most errors had to do with pronoun ambiguity and POS mistagging cases.

Chunking

We found no available chunking tool for French; we thus decided to turn to parsing for our syntactic feature, and to reduce the information as much as possible. Indeed, deep parsing is often regarded as too comprehensive for many NLP tasks, and as far as our linguistic features were concerned, we wished to keep track of the amount of information they would involve, to avoid overwhelming our models and creating unnecessary data sparsity.

The parser used was Arun and Keller's Collins-style lexicalised probabilistic context-free grammar parser (2005). A PCFG conditions the parse on the mother node by estimating the probability of its expansion from a manually annotated corpus. In a lexicalised PCFG, non-terminals are annotated with their lexical head. This parser was trained on Abeillé et al (2000) French Treebank corpus extracted from LeMonde journal. The syntactic tagset used involves the following:

SENT, VN, VPinf, NP, PP, AP, VPpart, PREF, Ssub, Srel, Sint, Adp.

Figure 6 : Tagset for French Parsing

VN stands for Verbal Nucleus, and involves in a flat structure the modifiers attributed to the head verb. VPpart is for verbs in the past or present participle form. VPinf opens an infinitival clause. AP is the Adjectival Phrase, which is generally included inside the Noun Phrase. PREF stands for verb-subject structures adjoined by a dash. The other tags are straightforward. To reduce the complexity inherent to parsing, we decided to assign each word to its closest phrasal category. This way, we could approximate the output of a regular chunker, which is non-recursive and thus phrase boundaries do not overlap.

English features

POS tagging

To POS tag the English corpus, we used the Brill tagger (Brill, 1995). This tagger is a transformation-based error-driven learning tool, and it conjoins abilities from both rule-based and statistical methods. It is based on the following principle: a small amount of unannotated text is primarily tagged by assigning each word its most probable tag as learnt from a corpus. Then, the tagged text is compared to the same text, but manually tagged; the tagger learns from its errors and generates an ordered list of transformation rules to adapt its output to true tag sequences. An example to such rules is stating that an item tagged as a verb must be transformed into a noun if preceded by a determiner. The process is then iterated. The Brill tagger is a state-of-the-art POS tagger, and has been reported to reach performances of around 97% for English. The version we used was trained on the Penn Treebank Corpus (Marcus et al, 1993), and comprehended 48 POS tags (including tags for punctuation and currency symbols).

CC, CD, DT, EX, FW, IN, JJ, JJR, JJS, LS, MD, NN, NNS, NNP, NNPS, PDT, POS, PP, PP$, RB, RBR, RBS, RP, SYM, TO, UH, VB, VBD, VBG, VBN, VBP, VBZ, WDT, WP, WRB, WP$, $, #, “, ”, (,), , , ., :

Figure 7 : Tagset for English POS tagging

Lemmatization and morphological analysis

Lemmatization and morphological analysis for English were performed using Carroll’s Morpha (Minnen, 2001): this partially statistically trained, rule-based morphological analyzer takes a word and its part-of-speech as input and maps the word to its lemma and inflectional information using a set of morphological rules and a list of exceptions. Each rule is implemented as a regular expression that, according to the word’s ending and POS, separates the word base from its inflection and reconstitutes the lemma out of the truncated word. The analyser cover the following English inflectional suffixes[4]:

-s plural of nouns, 3rd person singular present of verbs

-ed past tense

-en past participle

-ing progressive of verbs

In evaluation performed by the authors, the analyser performed successfully for above 99% of the test set.

The Penn Treebank tagset used for POS tagging covers these inflections quite well: indeed, noun plural is represented by the NNS tag, 3rd person singular present tense verbs by the VBZ tag, etc. We thus decided to use our POS information both for part-of-speech and for morphological features, while the lemma was extracted from Morpha’s output.

Chunking

As we mentioned earlier, chunking is sometimes preferred to full parsing for its achieving relatively fast and accurate results, and because it provides sufficient information for some NLP tasks. For the English chunking feature, we used YamCha (Kudoh, 2000), which is the system that performed best in Conll2000 shared task on chunk identification (above 93% F-Score for chunk recognition on test data of the same domain as the training). It uses a statistical method called Support Vector Machines which searches for the optimal parameters to categorize data into two classes. The models proposed by YamCha were trained on the

[4] As described in the README which comes with the software.

Conll2000 task training data for English, consisting of sections of the Wall Street Journal corpus POS-tagged with Brill and chunk-tagged using the IOB notation scheme (where I stand for Inside a chunk, B for Begin a chunk and O for Outside), which is convenient for using the chunk information as a feature per word.

3.2.1.2. Building linguistically augmented data

A linguistically-enriched parallel corpus had to be built for our models to be trained on. In this section, we shall briefly explain how the NLP tools described above were run on our data and how their output information was collected and regrouped into the intended format. Some additional data was formatted for building the language models of the factored models, and for tuning and testing the combined feature models. All formatting codes were implemented using Python.

The linguistically-enriched parallel corpus

The uppercased versions of the French and English corpora of the Europarl were used for linguistic annotation. Indeed, SMT systems generally use lowercased versions of the data in order to avoid building different models for a same word with different cases. For the purpose of automatic annotation, it is better to use uppercased text, for many systems take this factor into consideration to compute the best analysis.

Annotating the English side of the parallel corpus was quite straightforward: as mentioned earlier, the Brill tagger and the morphological analyzer Morpha did not need to be trained. The Brill tagger was first run on the data; its output was then fed into Morpha, which recognizes the Penn Treebank tagset. Morpha outputted the lemma, the morphological inflection of the word and its part-of-speech. To create a new linguistically-enriched English corpus, we brought together the lowercased version of the English corpus (for lowercased surface words) and Morpha's output (for word lemma and POS), and we tokenized the units we were interested in; these units were then combined, separated by a tube in the newly created corpus. This was performed one sentence at a time, thus preserving the original word order and sentence alignment. Our chunking tool, Yamcha, was first trained on a chunk-annotated corpus, and then run on a formatted version of the POS-tagged English corpus (one word per line, word and POS separated by a space, sentence boundaries represented by two lines). We then restored the structure of the original corpus by getting rid of words from Yamcha's output and realigning the chunk tags (in place of words) in the original sentences. We thus obtained a corpus of chunk tags: these were added to our main linguistically-enriched

corpus in the same fashion as other linguistic features had been added. We kept aside the chunk-only corpus, for we intended to expand it as well with phrase head and sentence subject information obtained with the Abney's Cass chunker (1995). This direction was unfortunately left aside eventually; for lack of adequate software availability, we did not manage to achieve the French counterpart to such a feature, which was a necessary condition for our translation models.

To annotate the French side of the corpus, we first formatted the uppercase version of the corpus in order to maximise the annotation tools' performance. We thus erased spaces before apostrophes for words like "l'évènement" where the l+apostrophe is a determiner. We also noticed recurrent typo mistakes for occurrences where the verb and the subject pronoun are inverted, as in "pourrions-nous": the hyphen was thus restored. Finally, some sentences were devoid of a full stop, which was a problem because the TreeTagger outputted one word per line, and we needed some kind of flag at sentence boundaries to restore original sentences after the linguistic processing stage. We thus inserted dummy markers accompanied with a dot at the end of each sentence, all of which would be erased once sentences would be brought back. The morphological analyser Flemm was consequently run on the tagger's output: its own output included the word, POS and morphological information. In the next formatting steps, the morphology was disambiguated in order to be left with a unique analysis, surface word forms were lowercased, the various features were ordered according to the agreed convention, and sentence structure from the original corpus was brought back by aligning all the feature blocks next to each other, and finally replacing our markers by new lines. For the chunking feature, we fed Arun and Keller's parser with the uppercased version of the corpus, which performs tokenisation and POS tagging of its own before parsing. We then simply assigned each word to its closest syntactic tag and wrote word|tag pairs into a new corpus file.

Training data for augmented language models

For our combined feature models, language models were trained on different versions of the main linguistically-enriched corpus, according to the experiment in question. For the factored models, separate files representing one factor at a time were created to train on factor-based language models.

Tuning and test data formatting

For the combined feature models, we processed the development data in order to translate from linguistically-enriched input, and to compare the output to a linguistically-enriched

reference translation. We thus formatted the development data in the same way that had been performed for the training corpus, and extracted the necessary features in accordance with the current experiment. Finally, the test data was formatted in the same way. The reference translation only was not modified: rather, surface words were extracted from the output translation and compared to the reference text.

3.2.2. Training the models

Translation models were trained using our two declared approaches, combined feature models and factored models, for French-to-English and English-to-French translations on a limited portion of the linguistically augmented parallel corpora (50,000 sentences), as well as on the whole corpus.

3.2.2.1. Combined feature models training

These models were trained on selected representations of the Europarl parallel corpus for French and English with equivalent combined features for each language. This was quite straightforward for all combined features besides where the French morphological feature was involved, in which cases the English POS feature was used as its equivalent in the parallel corpus.

First of all, word alignment (which in this case is combined features alignment) was applied in the same way as had been performed for our baselines: GIZA++ was trained on these combined features corpora to find the best alignment for tokens in sentence pairs. This kind of linguistically-informed alignment had been successfully experimented in previous work by Giménez and Màrquez (2005), who acknowledged the “natural trade-off between the use of data views [combined features] and data sparsity”, but claimed they disposed of enough data. In our case, it was unclear whether the data enrichment should help the alignment, or on the contrary complicate it. But we were mostly interested, given the data at hand, to verify if this kind of models was at all relevant to solve certain specific problems in translation output: the potential improvement in translation models adequacy that this alignment bore was enough of an incentive for us to try it, despite the possible sparsity problems that it may cause. Phrase alignment was then extracted, upon which translation tables were built.

Several trigram language models were then trained for French and English on corpora that included the various feature combinations experimented. For example, two language models, one for French and one for English, were trained on sequences of word+lemma combinations for these languages.

Finally, Minimum Error Rate Training was performed using formatted tuning files where both the input text and the reference text were represented as combined features. The original MERT code had to be slightly modified, for it used to retokenize input words and thus separate our features.

3.2.2.2. Factored models training

The word alignment for factored models was trained on surface words only, so that the translation tables used were the same as those used in our baselines. In future, these translation models shall be trained on multiple factors. For the purpose of this thesis, we concentrated mainly on the generation of factors.

Generation models represent the relation between different factors in the target language. Generation tables are trained on the factored version of the target language corpus: they list the probability of sequences of factors to generate sequences of other factors. For example, in a table that generates POS from surface words, we learn what is the most probable POS tag sequence given a sequence of words. This way, we build independent models of linguistic factors that can be used in conjunction. To build generation tables, we configured the learning algorithm to train on certain factors of the target language corpus.

Several language models were trained separately over versions of the entire annotated corpus that included one factor only. Given that these models were much more general than the ones we had built with combined features, we could easily build 7-gram models to represent part-of-speech or morphological information sequences, which we hoped would help representing phrasal and sentence gender/number agreement structures.

Finally, we ran a version of the MERT training that had been adapted to factored models and thus used Moses to perform decoding involving the defined models.

Shortcomings of the current factored models

It appeared that the chunk factor posed problem to the MERT training (and thus the decoding in general) of factored models. Given that this feature had been based on parsing output for French, the resulting ratio of possible tags per word was very big: for instance, the adverb “actuellement” may have been found in an adverbial phrase, in a verb phrase, adjectival phrase, etc… The problem with this fact is that the number of translation hypotheses explode while generating from word to chunk factor, and this probably caused the program to crash. Solutions envisaged include using a real chunker for French, as it is non-exhaustive and

outputs much simpler analysis, or finding a way to prune out less likely hypotheses generated by the decoder. But for now, we had to abandon experimenting with the chunk factor.

3.2.3. Testing the models

Decoding was performed using Pharaoh for the combined feature models, and Moses for the factored models.

3.2.3.1. Combined feature models

Appropriate versions of our linguistically enriched test data (i.e. including the appropriate combinations of features tested for that same experiment) were fed into Pharaoh's decoder, of which settings were the same as for our baselines (c.f. section on baseline decoding). Only the language models changed: for each experiment, the language model represented the combined features tested for the target language. The models tested were combinations of:

- word and POS
- word and morphology
- word, POS and morphology
- word and lemma

Each experiment was done in both translation directions, with models trained both on a limited section of the parallel corpus, and on the whole of it; in total, 16 experiments (8 for each translation direction) were performed.

3.2.3.2. Factored models

Decoding with Moses was performed on our regular word only test data. Moses' configuration differs from Pharaoh's in several aspects. First of all, it includes mapping steps for translation and generation. Translation from source to target factor is always performed first, then a generation step can be performed. Secondly, several language models are defined that represent the different factors involved in translation and generation. In our experiments, the models involved were thus a lexical phrase-based translation model, a generation model from lexical level to one or more other linguistic factors, and one language model for each target language factor (translated and generated). Other models that played a role in the translation process were the word penalty and distance-based reordering model. Probabilities from all models but the reordering model, are included in the future cost estimation to choose the most probable hypothesis. We did not test exactly the same models for French-to-English

and English-to-French translations, for POS was also considered as English's morphological factor.

The models tested for French-to-English translation were:

- Word factor translation, word to lemma generation
- Word factor translation, word to POS generation
- Word factor translation, word to POS and lemma generation

The models tested for English-to-French translation were[5]:

- Word factor translation, word to lemma generation
- Word factor translation, word to POS generation
- Word factor translation, word to POS and lemma generation
- Word factor translation, word to morphology generation
- Word factor translation, word to morphology and lemma generation
- Word factor translation, word to POS and morphology generation

Each experiment was performed with models trained both on a limited section of the parallel corpus, and on the whole of it; in total, 18 experiments (6 for French-to-English translation, 12 for English-to-French translation) were performed.

[5] The translation model for English-to-French translation with Moses was trained on a different corpus than the French language model; this fact affected the output results for these models, especially for words involving apostrophes which had been tokenized differently. Evaluation results for these experiments should thus be considered with this fact in mind.

Chapter IV: Results

4.1. BLEU scores

The BLEU scores obtained on our various experiments are presented in the following sections.

4.1.1. Combined Feature Models

The BLEU scores for combined feature models for French-to-English translations were lower than the baseline for models trained on a limited size corpus; for models trained on the whole corpus, two models scored above the baseline: combined word, POS and morphology features, and combined word and lemma features.

For English-to-French translations, the combined word and lemma model trained on a limited corpus size scored above the baseline; all other models (including those trained on the full size corpus) were below the corresponding baselines.

French-to-English	Trained on whole corpus	Trained on limited corpus
Baseline (Words)	30.26	**27.38**
Word + POS	29.97	26.16
Word + Morph	30.17	25.95
Word + POS + Morph	30.30	26.21
Word + Lemma	**30.36**	26.34

Table 21 : BLEU results for French-to-English translations with Combined Feature Models

English-to-French	Trained on whole corpus	Trained on limited corpus
Baseline (Words)	**32.42**	28.08
Word + POS	31.10	26.58
Word + Morph	31.07	26.73
Word + POS + Morph	31.02	26.96
Word + Lemma	32.18	**28.18**

Table 22 : BLEU results for English-to-French translations with Combined Feature Models

4.1.2. Factored Models

Factored models for French to English translations scored a little better than the baseline with word to POS generation trained on a limited corpus size. For models trained on the whole corpus, the word to lemma generation model was better than the baseline.

In English to French translations, the word to morphology model performed best, and the word to POS generation model performed better than the baseline.

French-to-English	Trained on whole corpus	Trained on limited corpus
Baseline (Words)	29.71	26.00
Word → POS	29.82	**26.02**
Word → POS, Lemma	29.61	25.98
Word → Lemma	**29.93**	25.67

Table 23 : BLEU results for French-to-English translations with Factored Models

English-to-French	Trained on whole corpus	Trained on limited corpus
Baseline (Words)	27.73	**21.86**
Word →POS	28.17	21.52
Word → POS, Lemma	25.17	21.68
Word → Morph, Lemma	27.54	21.44
Word → Morph	**28.25**	21.74
Word → POS, Morph	27.55	21.61
Word → Lemma	27.80	21.77

Table 24 : BLEU results for English-to-French translations with Factored Models

4.2. Manual evaluation

Manual evaluation was performed on 5 experiments. This evaluation did not aim at comparing the models in use (combined features and factored models); rather, it was meant to reach a better understanding to what extent the linguistic information in the models helped, and in what ways did they improve translation quality in each specific language, locally and globally. We analysed output from models trained on 50,000 sentences, as had been done for our baselines, in order to enable comparison.

With regard to the combined feature models, we evaluated the best performing experiment for English to French translation based on combined word and lemma features. It

performed a little better than the baseline. The two other combined feature models which were checked on manually were also English-to-French translations, which obtained a lower score than the baseline: combined features of word and morphology on the one hand, and of word and POS on the other. We worked with combinations of two features only, to have a better idea of their actual impact.

As to the factored models, we checked on two experiments translating from French to English: the word-to-POS generation experiment, which scored slightly better than the baseline on the BLEU scale, and word-to-POS-and-lemma generation, which scored a little below the baseline. Here, we had the opportunity of verifying the conjoined effects of independently influential factors on the output quality, according to BLEU score.

4.2.1. English to French translation, combined models

- Source words inserted in output: around 3.3% of the text were unknown words. This is almost twice as much as found in our Pharaoh baseline for English-to-French translations. The problem of sparse data creates an important disruption in the translation quality, and it is thus enhanced by this type of model.
- Translations which were completely wrong, and had no relation whatsoever with the source text, were handled about as well as the baseline: the range for the 3 models analysed was between 0.64% (for the word-lemma model) to 0.9%, while 0.77% of the text had been involved for the baseline. It may be argued that added POS and morphology information complicate the word alignment more than lemma information does, possibly because they create more tokens for a same word.
- The missing words phenomenon was also more prevalent in these two combined models: 0.23% and 0.35% of word-POS and word-morphology output were concerned, against 0.16% in the word-lemma model, which is close to what the baseline got. In general, many function words such as prepositions or determiners were missing, but sometimes, such was the case for nouns and verbs essential to the understanding of the sentence.
- The translation of English expressions and fixed structures into French was generally bad: while 46.84% of expressions had been wrongly translated in our baseline, 54% and 50.4% of these were literally and wrongly translated in word-lemma and word-morphology respectively. However, a decrease of this number was achieved by the word-POS model, with 43.23% of wrongly translated expressions. This finding may

well support the hypothesis that models informed with part-of-speech may better grasp word sequence structures, especially if this information is transferred (or translated) from source to target language. For example, “i can only agree with them” was translated into the fairly complex structure “je ne peux qu’être d’accord avec eux” (8), where the word “ne” generally brings on negation, while in this case it rightly conveys exclusiveness. Such was not the case in the baseline and the word-lemma models. However, complex structures which involved embedded clauses that had been wrongly translated by our baseline (for instance, by inserting a subject before and after the embedded clause) were not solved by either of these model. For instance, in “as a great many non-governmental organisations in the various countries of the european union have proposed” (23), the translations generally missed out the pronoun that should trace back to the subject in a correct French sentence (“comme un grand nombre d’organisations *l’ont* proposé”, i.e. “have proposed *it*”). Regarding subjunctive tense in subordinate clauses, our morphologically-enriched model did not seem to favour the right use of tense in context, although it is hard to be decisive on this issue for three such cases only were noted in the analysed text. The word-POS model did provide the right inflection in two of these cases: “so that we can” was translated “afin que nous puissions” (135) and “If the Union were to enlarge rapidly” (86) was translated “si l’union devait élargir rapidement”. Note that in the latter example, the French verb should be in its reflexive form “s’élargir”. Although reflexive expressions which exist in French but not in English were generally not grasped by our models, reflexive pronouns were mostly well translated, especially when preceded by a preposition (“in itself”, “for ourselves”), which translate into equivalent expressions in French (“en soi”, “pour nous-mêmes”). Finally, relative clauses like “the substances she referred to”, which in French require the insertion of a relative pronoun, were usually wrongly translated by all our models, with a slight tendency for the word-lemma model to do better than the others.

- The problem of wrong prepositions (i.e. right POS but wrong word) was much increased by all our models in comparison to our baseline. We did not expect this problem to be dealt with by our models.
- Wrong POS translation was best dealt with by our word-POS and word-morphology combined models, as hypothesized: the problem was reduced to 0.35% of the total number of output words, compared to 0.51% in our baseline. For instance, all of our models manage to disambiguate the verb “avoir” from its auxiliary and verb meanings,

so that phrases like “has to demonstrate” were translated as correct variants of “doit démontrer” (i.e. with the meaning “must” as opposed to “possesses”). Determiner/pronoun confusion, such as in “has underlined this once more” (translated by our baseline model “a souligné l’une fois de plus” or literally “has underlined the one more”) was not resolved by our models; this issue is complicated by the fact that reordering of the pronoun is also required before the verb in French. Finally, the models did not seem to influence the requirement of certain expressions to be followed by a certain part-of-speech in French.

- Tenses translation was slightly better for the word-POS combined feature model in comparison to the baseline (one mistranslation less than the baseline). The word-morphology model did about as well as the baseline. Given the many inflections in French, it may be that the sparse data problem is one cause to these results.
- Wrong agreements were not improved by our models: the word-lemma model performed about as well as our baseline, which had performed badly on 2.76% of verb and noun phrases conjoined. The word-POS model performed less well (2.84% agreement errors). Surprisingly, and in opposition to our hypotheses, the word-morphology model performed the worst (3.98% agreement errors). We noted that 6 of the 18 subject-verb agreement problems encountered in the baseline output had been resolved by the word-morphology model; however, various new problems were generated. Besides usual problems caused by the presence of embedded clauses, gender agreement was often not so well treated; in fact, our morphological feature did not include gender information for many nouns – only for those which had male and female possible inflections – and not at all for names (in French, proper nouns also bear a gender quality). The word-lemma model treatment of agreements was found to be very similar to our baseline’s, with a few positive differences.
- Ordering errors in noun phrases including modifiers (adjectives and determiners) were much more frequently generated by our combined feature models than they had been in our baseline for English-to-French translation (with 2.6% more ordering errors on NPs than for our baseline). This opposed our previous hypothesis that POS information may help bias towards the right ordering of components in a phrase. Errors occurred mostly in coordinated compound noun phrases. Ordering of linguistic components on the sentence level was best managed by the morphologically-enriched model, with 0.9% errors of the total clauses in the text, compared to our baseline’s

3.1% errors. The POS-informed model performed about half the syntactic reordering errors of the baseline.

4.2.2. French to English translation, factored models

- Source words inserted in output: around 2% of the text were unknown words for our baseline model and the two factored models. This is expected, as the word alignment was the same for all models. It shows that the factored models do not enhance sparse data problems.
- Translations with no relation whatsoever with the source text were less frequent in our factored models: they represented 0.85% of Moses baseline output, and 0.69% and 0.59% for word to POS and word to POS-lemma generation models respectively.
- The missing words phenomenon was also less prevalent in the two factored models: 0.56% and 0.51% of word to POS and word to POS-lemma models output were concerned, against 0.8% in the baseline.
- The translation of French expressions and fixed structures into English were overall good: the word to POS and lemma generation model made mistakes on 15.3% of total expressions, while the baseline committed 16.26% errors. The word to POS generation model was a little worse on this issue (one more error was found). For instance “que le parlement ne présente plus” (25) was rightly translated as “that parliament no longer presents” by our word to POS generation model, instead of the baseline’s version “parliament presents more”. “avec beaucoup de générosité” was rightly translated “with great generosity” (29). Nevertheless, neither of the models managed to solve the problem of relative clauses lack of correspondence between French and English (we noted sentences such as “major change […] which Europe must adapt”, missing out the final preposition “to”). Reflexive verbs were also dealt with incorrectly, especially when embedded clauses were involved, as in “nous puissions, […], nous fixer” (39), which means “we could set ourselves”, and was translated by all models as “we can, […], we have to”.
- Wrong prepositions were slightly more common in our models in comparison to our baseline.
- Wrong POS translation rate was decreased by 0.1% with the word to POS generation model, which is relevant given that this problem represented only 0.33% of our

baseline output translation. For instance, this model disambiguated successfully a determiners from a pronoun in “je ne peux que les rejoindre” (8) rightly translated “i can only join them” (while in the baseline, it was translated “I can only the join”).

- Regarding tense errors, both our factored models performed as well as the baseline (2.99% errors). The POS information (which also took on the role of morphological factor) did not, therefore, seem to influence the output.
- Wrong agreements: verb-subject agreement was not influenced either by the use of a POS factor and the error ratio remained identical to our baseline’s.
- Ordering errors in noun phrases including modifiers (adjectives and determiners) were about the same as the baseline; the word to POS generation model performed a little worse. Ordering of linguistic components on the sentence level was best managed by the POS-enriched model, with 1.49% errors of the total clauses in the text, compared to our baseline’s 4.19% errors.

4.2.3. Summary

In the light of our manual evaluation, it appeared that the word-lemma combined model was the closest to our baseline performance exactly because it is the one that causes the least modification to it; indeed, while POS and morphological information introduce more sparse data by assigning different analyses to a same token (thus creating multiple tokens), the lemma information probably had a limited such effect. The positive impact of the lemma feature represented by the BLEU score was thus looked at with some reservation. However, findings on the lemma information as part of the factored models for French to English translations supported the idea that this linguistic level does have a role in the reduction of translation errors, and helps the decoder to choose adequate terminology.

The word-POS combined and word to POS generation models appeared to influence to a limited extent the output’s fluency, by favouring grammatical word sequences in the target language, as was hypothesized; however, they were fruitless in dealing with complex compound noun phrases including adjectives and coordination. In combined feature models, the POS feature seemed to create more sparse data and to have worsened the word alignment. In both combined and factored models, the POS feature/factor influenced in a limited way POS disambiguation. However, in neither models did the POS feature/factor influence positively word agreement and verb tense, in opposition to our hypotheses. The conjunction

of POS and lemma factors was overall beneficial in the issues where the POS factor alone already performed well.

The morphological feature for translation into French as it was used in our models performed in a similar way as the POS feature, but it often gave worse results than the latter.

4.3. Fine-grained evaluation

In our various baseline evaluations, we had acknowledged difficulties of translation in the face of unseen words in the training data, part-of-speech ambiguity, and syntactic agreement. Our automatic error report achieved three things. The first was to estimate the number of source language words inserted in the translation output. The second was to give an overall estimate of the translation performance in comparison to the reference text with regard to the amount of untranslated words, and of words of which the lemma, but not the inflection, had been rightly translated. The third allowed us to obtain a general overview of how well our models performed with regard to the two other problems noted above: it gave us some information on how well the various POS categories in the reference translation were represented in the output, and on how frequently translated words were in their correct inflectional forms (thus giving, by extension, an understanding on subject-verb and noun-modifiers agreements). These aspects shall be reviewed now for both tested models.

4.3.1. Combined feature models

4.3.1.1. Source language words inserted in target output

The phenomenon of source words inserted in the target output was found to increase in importance when combined features were used, compared to our baseline. For French to English translations, the percentage of source words inserted in the output was around 4% for models trained on the whole corpus, between 6.1% and 6.6% for models trained on a limited corpus size. For English to French translations, between 3.2% and 4.2% for whole corpus trained models, between 4.7%-6.1% for models trained on a limited corpus size. However, it made no real difference if more than 2 features were combined. The only model that limited this phenomenon and was closer to (but not as good as) the baseline from that point of view was the word-lemma model. For all models, at least a good half of the word types which were not translated had a morphological variant in the training data. For a detailed presentation of results, please refer to the appendix.

4.3.1.2. Mistranslation rate

The tables below represent the percentages of mistranslation (i.e. word mismatch between output and reference translations) and of mistranslations that did include the right lemma, out of the total number of words in the reference translation. We can observe that all models were improved by the addition of training data. For French-to-English translation, all combined feature models performed a little worse than the baseline when trained on a limited corpus size, a little better when trained on the whole corpus. The percentage of right lemma translation was steady (between 2.6-7% of the text). For English-to-French translations, all our models performed less well than our baselines, except the word-lemma combined model trained on a limited corpus size, which performed about as well as its corresponding baseline. The percentage of right lemma translation was ranged between 10.6-11.6%, which represents an important portion of the translation.

French-to-English	Trained on whole corpus		Trained on limited corpus	
	Mistranslation rate	Right lemma translation rate	Mistranslation rate	Right lemma translation rate
Baseline (Words)	36.1	2.7	38.9	2.6
Word + POS	36	2.7	39.2	2.6
Word + Morph	35.9	2.6	39.2	2.6
Word + POS + Morph	35.9	2.7	39.1	2.6
Word + Lemma	35.9	2.7	39.1	2.6

Table 25 : Mistranslation evaluation for French-to-English translations with Combined Feature Models

English-to-French	Trained on whole corpus		Trained on limited corpus	
	Mistranslation rate	Right lemma translation rate	Mistranslation rate	Right lemma translation rate
Baseline (Words)	34.1	11.3	37.2	11.6
Word + POS	34.8	10.7	38.3	10.8
Word + Morph	34.8	10.6	38.4	10.8
Word + POS + Morph	34.8	10.7	38.3	10.7
Word + Lemma	34.3	10.8	37.2	11

Table 26 : Mistranslation evaluation for English-to-French translations with Combined Feature Models

4.3.1.3. Mistranslation by POS

There were two cases of wrong POS translation found in our baselines: cases where the decoder chose completely wrong translation because of POS ambiguity, and cases where the decoder picked a word from the right family, but assigned it the wrong syntactic role. With our fine-grained evaluation, the first case is represented by the overall error rate performed for each POS category translation; the second is represented by our classified right lemma translation with wrong inflection. We were also interested in finding out about possible wrong agreements, which were also reflected by the percentage of right lemma translation with wrong morphology. From our baseline evaluation, it came out that POS categories that were most concerned with wrong translation were verbs, function words such as adverbs and prepositions, and adjectives and nouns. In translation to French, pronouns were also part of this list.

The models which we had hypothesized to improve verb tense and subject-verb agreement mistranslations involved POS or/and morphological features. It appears from our findings that word-POS, word-morphology and word-POS-morphology models all performed 1% more translation error with verbs than corresponding baselines for English-to-French translations. For French-to-English translations, models generally performed about as well as the baseline; the only small noticeable improvement was for infinitival verbs, which in the word-morphology model underwent 0.1% improvement for the model trained on the limited corpus, and 0.4% for the model trained on the whole corpus, and in the word-POS-morphology model underwent a 0.3% improvement trained on limited data, and 0.6% trained on the whole data.

Noun-modifier agreement was also hypothesized to be positively influenced by these models. Such was not the case, according to our findings, for English-to-French translation. For French-to-English translation, a little 0.1% improvement was achieved with the word-lemma model trained on the whole corpus. Neither did these models get more noun lemmas right than our baseline did. Another 0.1% improvement in nouns was achieved with the word-morphology model trained on the whole corpus.

4.3.2. Factored models

4.3.2.1. Source language words inserted in output

For factored models, the proportion of source words inserted in the output was left almost unchanged by the addition of factors in the model. For French to English translation, 3.9% of

the text were estimated to be unseen words for models trained on the whole corpus. For models trained on a limited portion of the data, this number was around 6%. For English to French translations, models trained on the whole corpus bore around 3.3% unseen words; models trained on the limited size corpus bore around 4.8% unseen words. For French to English translation, the number of mistranslated word types which had a morphological variant in the training data was at least half the types found. For English to French translation, this number was higher still: around 65.6% of mistranslated words had the right lemma for models trained on a limited corpus, while for models trained on the whole corpus, this was the case for around 84% of the text.

4.3.2.2. Mistranslation rate

The numbers in the tables below are percentages of mistranslation and of mistranslations that did include the right lemma, out of the total number of words in the reference translation. Besides the fact that all models were improved by the addition of training data, it is interesting to note that the percentage of mistranslations with a correct lemma was also influenced by the training data size, and even more so for English-to-French translations.

For French-to-English translation, the best performing models were word to lemma and word to lemma and POS generation when trained on a limited corpus size. When trained on the whole corpus, only the word to lemma and POS generation model performed better than the baseline. The percentage of right lemma translation tended to grow a little for models trained on the whole corpus, and ranged between 2.6-8% of the text. For English-to-French translations, the word to morphology and lemma generation model performed a little better than the baseline when trained on a limited corpus size. When trained on whole corpus, the best performing models were word → POS, lemma and word → lemma, which performed only 33.9% and 34.8% errors respectively. Increase in the percentage of right lemma translation was observed for all other models (increase of between 0.7% to 1.6%); for our best performing models, a decrease in this number was also rewarded by a decrease in general mistranslation. The percentage of right lemma translation was ranging between 9.5-11.1% of the text.

French-to-English	Trained on whole corpus		Trained on limited corpus	
	Mistranslation rate	Right lemma translation rate	Mistranslation rate	Right lemma translation rate
Baseline (Words)	36.2	2.7	39.3	2.7
Word → POS	36.1	2.7	39.2	2.6
Word → POS, Lemma	36.6	2.6	39.4	2.6
Word → Lemma	36.2	2.8	39.7	2.7

Table 27 : Mistranslation evaluation for French-to-English translations with Factored Models

English-to-French	Trained on whole corpus		Trained on limited corpus	
	Mistranslation rate	Right lemma translation rate	Mistranslation rate	Right lemma translation rate
Baseline (Words)	35	10.6	37.4	9.6
Word → POS	35.2	10.9	38.2	9.6
Word → POS, Lemma	33.9	9.6	37.4	9.8
Word → Morph, Lemma	35.5	10.3	37	9.6
Word → Morph	35.3	11	38	9.5
Word → POS, Morph	36.3	11.1	38.1	9.5
Word → Lemma	34.8	10.7	38.6	11.1

Table 28 : Mistranslation evaluation for English-to-French translations with Factored Models

4.3.2.3. Mistranslation by POS

First of all, we wished to check how our POS and morphology factors influenced verb tense and subject-verb agreement mistranslations. It appeared that the word to morphology generation model trained on the whole corpus for English-to-French translation improved verb translation of 0.1%, and when the lemma factor was added, 0.5% improvement was achieved. 0.9% improvement in verb translation was obtained with the word to POS and lemma generation models trained on the limited and on the whole corpus. In French-to-English translation, verb translation was improved by 0.4%, and noun translation by 0.2% for

the word to POS model trained on the limited corpus size. When the lemma factor was added, noun translation was improved by 0.5%. The combination of the lemma factor and higher level linguistic representations of the word thus seemed beneficial in this regard. Also, all models involving the POS factor (besides the word to POS and morphology generation model, which achieved poor general results) achieved a good reduction of translation errors on determiners for English to French translation.

Noun translation was barely improved by the word to lemma and word to morphology and lemma generation models, with a decrease of 0.5% in noun mistranslation. This hardly supports our hypothesis that models enriched with lemma information would bias the decoder to choose words of the right family. Generally, the lemma factored models did not show much improvement in translating the right lemmas: the fact that the word to lemma generation model trained for English-to-French translation on the limited size corpus achieved 1.4% improvement in translating adverbs, 8.1% in translating determiners and 3.7% in translating adjective lemma should not mistaken us, for we need to keep in mind that this model got the worst overall mistranslation score in the factored models for English-to-French translation, probably thus influencing these scores. Otherwise, it did about as well, or worse than the baseline in recognizing lemmas.

4.3.3. Summary

The study of unseen words inserted in the output translation in their original form showed us that factored models were indeed better to deal with this problem, as they did not create extra sparsity of data. This was true especially for English to French translations, where the higher complexity of French morphology apparently increased ambiguity and sparseness of data for our combined feature models.

With regard to mistranslation rate, combined feature models performed better than factored models. This does not univocally determine that combined feature models performed better than factored models, as these numbers do not take into account word insertion and ordering. However, this might tell us something about the necessity of translating features/factors from source to target language. The models which performed the least translation errors were combined word and lemma feature models trained on the whole corpus for French to English translations. For factored models, the best performing model for French-to-English translation from the viewpoint of mistranslation was the word to POS generation model (trained both on limited and extended data); for English to French translation, the best performing model was word to POS and lemma generation.

The evaluation of error distribution by POS showed very little improvements made by either one of our models on specific translation problems such as word inflection and correct word choice.

Chapter V: Conclusions and Future Work

In this thesis, we investigated whether linguistically-enriched phrase-based statistical machine translation models could improve translation output quality from the viewpoints of adequacy (i.e. conserving the original text meaning) and fluency (i.e. grammaticality), by providing additional information for a word and its surrounding words.

To do so, we first engaged in different types of output evaluation that were meant to dig into the problem of phrase-based translation from different angles. We believed that this step was necessary to better understand where problems stood in the existing machine translation models, and to better design ways of resolving them. Our methods were extensive, including manual evaluation as well as two automatic evaluation methods. One difficulty in generated language evaluation is to clearly define what is being looked for. While the BLEU metric counts ngram matches and the Word Error Rate computes word insertion and deletion, other methods may take into account the use of synonyms. In our sense, refining these methods by conjoining automatic and manual evaluations could be found rewarding: for instance, it may be of use to manually analyse errors found automatically, and to adapt automatic methods accordingly. We found it difficult to decide, for example, whether nouns determined as mistranslated by our automatic method were real translation mistakes, or if the alternative offered by the decoder could be seen as acceptable. Also, following our analysis of our automatic evaluation method, it appeared that words of a same family had been translated which were not accounted for by our method, because lemmatization is more restricting than stemming from this point of view; it may be interesting to include stemming in our calculations next time. Evaluation of language generation systems is a wide and important area of research, which still needs to propose innovative solutions.

The two models proposed to augment translation models with linguistic information posed both problems of their own: while the combined feature models enhanced sparse data problems, the factored models as they were supported by Moses at the time of this research could only efficiently generate factored target linguistic levels. We thus took on to check both models and see if, on the one hand, source to target feature translation (i.e. using combined feature models), and on the other, target factor generation (i.e. using factored models), could improve translation quality. We designed a parallel corpus enriched with linguistic features/factors to be trained on by both models.

One issue we wished to investigate was the effect of unseen data on our models, given that large parallel corpora are not necessarily easy to acquire for languages other than the main European ones, and especially given the fact that Pharaoh and Moses decoders do not implement smoothing or back-off methods for these occurrences. The combined feature models were much less robust to data sparsity than the factored models. Indeed, combined features allowed less flexibility with the models, requiring the use of surface words as a necessary feature in the models. Training models independently from one another, as this is the case with factored models, avoids enhancing the sparse data problem, and also allows the flexible modelling of linguistic feature sequences. An independent sequence model for POS information may be much more informative on a language's behaviour than a sequence model where surface word and POS are attached to each other, thus missing the effect of generalization that this deeper linguistic level can provide. Although the use of more than 2 features was not found to further worsen the models, the sparse data problem inherent to the combined feature models complicated the word alignment. For factored models, it is possible to use surface word only alignment, which was also the one used for our baselines. Nevertheless, we should definitely look into bettering the word alignment, given the many problems currently generated by all of our models because of sparse data and alignment errors. There would also be some interest in using some backoff method to more general linguistic levels such as lemma and POS. According to our findings, most unseen words which were inserted in their original form into the output translation by current translation models, could have benefited from being recognized as family-related to other words. With such information, we may implement a backoff system as the one proposed by Kirchhoff and Yang (2006), which would greatly reduce the impact of the sparse data problem on output translation.

The other essential question we wished to inquire was which features could help improve our models, what would be their impact on the output, and how would they behave separately and in conjunction, and within the different modelling frameworks proposed. In short, what are the optimal features for translation. The answer to this question is not straightforward when looking at our experimental results. Our main hypotheses on word agreement and verb tenses were not supported by our findings. This may have to do with the nature itself of our morphological feature; indeed, our morphological feature for French provided a word's gender identity only if it had a possible opposite gender counterpart, thus undermining many attempts for gender agreement. Another example of its limits is that it did not differentiate between personal and reflexive pronouns. There would thus be a need for

more efficient morphological annotation. Finally, we did not get to measure in detail how the word to morphology generation model for English to French translation trained on the whole corpus achieved the best BLEU score for that set of experiments. We had hypothesized that factored models could possibly generate right French inflectional forms in context with generation from word to morphology, using a language model that could possibly bias the decoder to choose right agreements. It would be interesting to verify if factored models managed to generate agreements for French. For translation into English with factored models (where the POS was the morphological feature), such was not the case.

The modelled features/factors which were found to provide limited improvement on several issues such as non-literal translation of fixed and idiomatic expressions, overall syntactic ordering and POS disambiguation were the lemma and the POS information. Word agreement, verb tenses and complex structures such as coordinated phrases, embedded clauses and compound nouns were not affected by the use of these linguistically-enriched models. In particular, embedded clauses created several agreement errors, as the language model cannot handle long-distance dependencies. We may suggest at this point to try different features in future work. Indeed, we did not manage to use our chunking feature, which had a strong hypothesized role in helping resolve word inflection errors. A chunking feature may replace words in their syntactic context, even more so if it is conjoined with POS and lemma information, to possibly achieve grammatical agreement. Also, a syntactic model including head dependencies may help recognize a sentence main arguments and the structure of word agreement. However, before we think of enriching our features further, the space explosion problem we had encountered with factored models, when trying to decode with our chunk feature, should be resolved.

We shall now conclude on the models' performance themselves. First of all, with regard to the specific languages experimented, English to French translation was generally found to be harder than the other way round: the richer morphological diversity of French was assumed to be one part of the issue (i.e. non-corresponding tenses and gender features), while structural complexity of high level writing was another. In general, translation to English for the models that were manually evaluated was always more easily comprehensible than to French. Secondly, with regard to the actual models in use, the combined feature models had the strong disadvantage, on the one hand, to widely depend on the occurrence of expressions in the training data: by disambiguating translation possibilities, it also restricted the correct translation possibilities for the alignment algorithm. But what we are really looking towards is the generalization of models, not the contrary. On the other hand, it was noted that translation

of features from source to target language was probably an important step to representing the linguistic knowledge in the output translation: indeed, our combined feature models performed less mistranslation errors than the factored models. We are thus looking forward to the possibility of performing factored translation including translation steps from source to target factor.

On another level, it appeared from our findings that the POS feature/factor has a positive influence on overall syntactic order; it would be very interesting to consider adding such linguistic features/factors to the reordering model. Finally, the language models used for the factored models were not “factored” language models per se; we suggest for future work to integrate factored models with a possible back-off to more general representations of a word to deal with unseen linguistic occurrences. This way, we may exploit those different linguistic levels independently, while concentrating their information.

References

A. Abeillé, L. Clément, A. Kinyon (2000). Building a treebank for French. In *Proceedings of the LREC 2000,* Athens, Greece.

S. Abney(1995). Partial Parsing via Finite-State Cascades. In *Natural Language Engineering 1*, Cambridge University Press.

A. Arun, F. Keller (2005). Lexicalisation in Crosslinguistic Probabilistic Parsing: The Case of French. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 306-313. Ann Arbor, MI.

E. Brill (1995). Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. Computational Linguistics, vol. 21, no. 4, 543-566

F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer (1993). Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, vol. 19, no. 2, 263-311.

C. Callison-Burch, M. Osborne, P. Koehn (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of EACL-2006*.

D. Chiang (2005). A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistic*.

S. Corston-Oliver and M. Gamon (2004). Normalizing German and English Inflectional Morphology to Improve Statistical Word Alignment. In *Lecture Notes In Computer Science*, ISSU 3265, pp. 48-57.

N. Fiammetta, J. Christian (2000). Flemm: un analyseur flexionnel du français à base de règles. Traitement automatique des langues pour la recherche d'information, vol. 41(2), pp. 523-547.

J. Gimenez and L. Marquez (2005). Combining Linguistic Data Views for Phrase-Based SMT. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pp. 145–148, Ann Arbor.

H. Hoang (2006). Moses, a Phrase-Based Multi-Factor Decoder for Machine Translation. Developer's Manual, unpublished.

K. Kirchhoff and M. Yang (2005). Improved Language Modelling for Statistical Machine Translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, 125–128, Ann Arbor.

K. Kirchhoff and M. Yang (2006). Phrase-Based Backoff Models for Machine Translation of Highly Inflected Languages. In *Proceedings of EACL*.

K. Knight and D. Marcu (2005). Machine Translation in the year 2004. In *Proceedings of ICASSP*.

K. Knight, K. Yamada (2001). A syntax-based Statistical Translation Model. In *Proceedings of ACL*, 523-530, Toulouse, France.

P. Koehn (2005). Europarl: A Multilingual Corpus for Evaluation of Machine Translation. *MT Summit 2005*.

P. Koehn, C. Monz (2006). Manual and Automatic Evaluation of Machine Translation between European Languages. *NAACL 2006 Workshop on Statistical Machine Translation*.

P. Koehn (2004). Pharaoh, a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models, User Manual and Description for Version 1.2. USC Information Sciences Institute.

P. Koehn, F.J. Och, D. Marcu (2003). Statistical Phrase-Based Translation. In *Proceedings of HTL-NAACL 2003*, 48-54, Edmonton.

T. Kudoh, Y. Matsumoto (2000), Use of Support Vector Learning for Chunk Identification. In *Proceedings of CoNLL-2000 and LLL-2000*, Lisbon, Portugal.

M. Marcus, B. Santorini, M. Marcinkiewicz (1993). Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, 19(2).

G. Minnen, J. Carroll and D. Pearce (2001). Applied morphological processing of English. *Natural Language Engineering*, 7(3), 207-223.

S. Nießen and H. Ney, (2001). Toward hierarchical models for statistical machine translation of inflected languages. In *Proceedings of the Workshop on Data-Driven Methods in Machine Translation - Volume 14* (Toulouse, France, July 07 - 07, 2001). Annual Meeting of the ACL.

F.J. Och (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics.*

F.J. Och, C. Tillmann, H. Ney (1999). Improved Alignment Models for Statistical Machine Translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pp. 163-166.

F.J. Och, H. Ney (2000). Improved Statistical Alignment Models. In *Proceedings of ACL 38*.

K. Papineni, S. Roukos, T. Ward, W.J. Zhu (2001). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

G. Pilard (ed.), (2000). Harrap's Shorter French and English Dictionary, Chambers Harrap Publishers Ltd, Edinburgh, UK.

M. Popovic, A. de Gispert, D. Gupta, P. Lambert, H. Ney, J.B. Marino, M. Federico, R. Banchs (2006). Morpho-syntactic Information for Automatic Error Analysis of Statistical Machine Translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pp. 1–6, Association for Computational Linguistics, New York.

H. Schmid (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*.

A. Stein, H. Schmid (1995). Etiquetage morphologique de textes français avec un arbre de décision. Traitements probabilistes et corpus, vol. 36, 1-2, pp. 23-35

A. Stolcke (2002). SRILM – An Extensible Language Modeling Toolkit. Proceedings of ICSLP.

J. Véronis (coordinator) (1996). Multext, Multilingual Text Tools and Corpora. In http://aune.lpl.univ-aix.fr/projects/multext/, Centre National de la Recherche Scientifique.

APPENDIXES

A. Manual evaluation of Pharaoh baselines[6]

1. English to French Translation

a) Source language words inserted in the translation

Number of occurrences: 79

Occurrences:
appealing, oneself, weigh, understating,environmentally-friendly, all-embracing, enforceable, summarises, differentiated, paces, duck, meteoric, dispel, aforementioned, ammunition, allotted, decentralised, pessimist, halves, allegiance, thrifty, sore, topped, clever, steady, warrant, hijacked, anti-nuclear, malta, blinded, transposed, daddy, egoism, populism, overwhelmed, feathers, materialism, jeopardised, ordered, orphaned, insight, nineties, industrious, rally, assiduous, long-overdue, poisonous, long-lived, dioxin, aflotoxin, mercury, notify, emphasised, non-marine, misguided, aggregated, lore, commodities, merge, scattered, percentages, feedingstuff, feedingstuffs, pledged, coumpound, feedingstuffs, ingredient, feedingstuffs, episodes, progressive, inception, withstand, headlong, aggravate, sidelong, shifts, populist, fascistic, appearing

b) Wrong word: Translation not understandable without looking at source text

Number of occurrences: 17

1. “a clear idea of the rights” = “une idée claire de l'homme” (5)
2. “12.30 p.m” = “12.30 30” (11)
3. “into the treaty at the nice summit” = “dans le traité fasse lors du sommet” (20)
4. “the countries which joined later” = “les pays qui depuis plus tard” (32)
5. “of catching up with” = “de merlu de” (32)
6. “will also be governed” = “seront également besoin” (34)
7. “the task of acting as rapporteur” = “la tâche de faire en tant que rapporteur” (50)
8. “a key role will fall” = “un rôle clé à baisser” (65)
9. “are closer together” = “sont plus défavorisés” (67)
10. “inhuman transport of live animals” = “le transport d'animaux longtemps ces procédés” (72)
11. “we are moving to a requirement” = “nous sommes d'une obligation” (73)
12. “off the roads” = “à l'orateur les routes” (76)
13. “property” = “de propriété intellectuelle” (80)

[6] The numbers in between brackets are line numbers in the translated text.

14. “which can build up” = “qui peuvent mettre en place” (114)
15. “we are beginning to hear” = “il s'entendre” (147)
16. “to bring peoples closer together” = “d'un des plus défavorisés” (148)
17. “unrestrained” = “sans aussi” (149)

c) Wrong word/phrase: translation has something in common with original text

Literal translation: Words translated as in source text

Number of occurrences: 52

Various fixed expressions

Number of occurrences: 34

Occurrences:

1. “the need to guard” = “la nécessité d'en garde” (6)
2. “i can only agree with them” = “je peux seulement d'accord avec eux” (9)
3. “less fuel” = “moins carburants” (16)
4. “such buses” = “telles les bus” (16)
5. “in favour of incorporating” = “en faveur d'inclure” (20)
6. “to enable work to continue on it “ = “pour permettre de continuer à ce travail” (23)
7. “as a great many non-governmental organisations in the various countries of the european union have proposed” = “comme un grand nombre d'organisations [...] ont proposé” (23)
8. “a general report at all” = “un rapport général à tous” (25)
9. “or so i hope” = “ou si j'espère” (26)
10. “this does not , of course , mean” = “ce n'est pas, bien sûr, signifie” (27)
11. “where major , long-term investments are needed” = “les grands, des investissements à long terme sont nécessaires” (33)
12. “complying with” = “exécuter avec” (37)
13. “in order for us to be” = “afin de nous être” (39)
14. “as indeed i was earlier” = “j'ai précédemment” (43)
15. “both extensive and decentralised” = “deux détaillée et decentralised” (44)
16. “to get my point across” = “obtenir mon point de” (45)
17. “we must make ourselves clearly understood” = “nous devons faire nous-mêmes clairement compris” (45)
18. “i see no need” = “je ne vois pas besoin” (46)
19. “which is also why” = “qui est également pourquoi” (48)
20. “will now speak” = “seront maintenant la parole” (49)
21. “can you have” = “ne peut que vous avez” (52)
22. “direct foreign investment” = “investissements étrangers directs” (no determiner) (57)
23. “we must use waterways” = “nous devons utiliser voies d'eau” (no determiner) (76)
24. “are already calling” = “sont déjà réclame” (83)
25. “as was stated in the eldr group” = “comme l'a indiqué dans la position du groupe eldr” (86)

26. “the council is not applying” = “le conseil n'est pas l'application” (96)
27. “more than anything” = “plus qu'il est quelque chose” (98)
28. “has turned out to be” = “a été d'être” (106)
29. “many of my fellow members” = “de nombreux députés de mon celle” (109)
30. there must at least be a requirement” = “il faut au moins être une obligation” (116)
31. “primary feed materials” = “matières alimentation primaire” (126)
32. “all due to” = “tout en raison de” (136)
33. “contaminants” = “contaminants” (137)
34. “following the pace” = “après le rythme” (146)

Tense in subordinate clauses

Number of occurrences: 3

1. “If the Union were to enlarge rapidly” => “si l’union étaient d’élargir” (86)
2. “so that we can” => “afin que nous pouvons” (135)
3. “it is important that this parliament sends” = “il est important que ce parlement transmet” (147)

Wrong use of subordinate clause

Number of occurrences: 5

1. “aims we all set ourselves” = “objectifs nous étions fixés” (9)
2. “what it is that is important” = “c'est ce qu'il est important” (92)
3. “those problems it is facing” = “les problèmes qu'elle est confrontée” (97)
4. “an argument we can use” = “un argument nous pouvons utiliser” (101)
5. “the substances she referred” = “les substances qu'elle a fait référence” (122)

Wrong use of negative form

Number of occurrences: 10
Occurrences:

1. ”does in any case” = “n'est en tout cas” (3)
2. “i have no problem saying it” = “j'ai aucun problème à le dire” (27)
3. “there can be no transitional period” = “il peut y avoir aucune période transitoire” (37)
4. “the report which it is my honour” = “le rapport qu'il n'est mon honneur” (64)
5. “which is why” = “qui n'est pourquoi” (67)
6. “which does justice to the challenge” = “qui n'est le défi de la justice” (67)
7. “can only dream of” = “ne peuvent avoir un rêve” (71)
8. “nor , certainly , has the falling rate of the euro helped” = “pas, certes, a la baisse des taux de l'euro aidé” (77)
9. “ people do , however , rally” = “cependant, les gens ne rally” (102)
10. “was there with me” = “il n'a été avec moi” (106)

Same POS but wrong word: wrong preposition

Number of occurrences: 13
Occurrences:

1. “, messages of concern at the economic and social problems” = “les déclarations d'inquiétude de problèmes sociaux” (2)
2. “weigh them up , for political discourse must always be realistic” = “weigh jusqu'à leurs discours politiques doivent toujours êtres réalistes” (9)
3. “i will vote to approve” = “je voterai d'approuver” (9)
4. “from greece to portugal” = “de la grèce pour le portugal” (14)
5. “pressure for adaptation” = “pression de l'adaptation” (30)
6. “since my first speech was three minutes” = “depuis ma première intervention a été allotted sous trois minutes” (43)
7. “for where are the millions” = “d'où sont les millions” (82)
8. “make up for lost time” = “qu'il fallait faire jusqu'à temps perdu” (100)
9. “to limit the transitional periods in the environmental sphere to a maximum of five years” = “limiter les périodes transitoires [...] d'un maximum de cinq ans” (103)
10. “on more than one occasion” = “sur plus d'une fois” (106)
11. “been reduced to production units” = “ont été réduits d'unités de production” (123)
12. “can be done in a couple of days” = “peut se faire dans deux jours” (133)
13. “is important to old as well as new sectors” = “est important d'anciennes ainsi que de nouveaux secteurs” (69)

Wrong POS

Number of occurrences: 23

Have – auxiliary or verb

Number of occurrences: 4

Occurrences:

1. “will not have to be held” = “n'aura pas d'être menées” (42)
2. “has to demonstrate” = “a de démontrer” (74)
3. “it would have to decentralise” = “il aurait d' decentralise” (86)
4. “farmers quite simply have to know” = “les agriculteurs ont tout simplement plus de savoir” (134)

Determiner/pronoun confusion

Number of occurrences: 3
Occurrences:

1. “.according to those in favour” = “selon ces en faveur” (16)
2. “has underlined this once more” = “a souligné l'une fois de plus” (37)
3. “draw from it sometimes” = “peut tirer de l' parfois” (145)

Wrong POS

Number of occurrences: 16

Occurrences:

1. “he also calls for benchmarking to spread best practice across the eu” = “il exige de meilleure pratique pour d'extension de l'ue” (17)
2. “completely legally enforceable rights” = “totalement juridiquement enforceable” (18)
3. “in nice” = “de bonne” (26)
4. “this results” = “ce résultat” (32)
5. “i cannot say that today of any country” = “je ne peux pas dire qu'aujourd'hui d'un pays” (47)
6. “specific mention” = “mentionner spécifique” (60)
7. “we all stand to benefit from enlargement” = “nous sommes tous d'avantage de l'élargissement” (75)
8. “and the obvious , but not public resistance” = “et l'évidente, mais pas l'opposition” (81)
9. “the vast majority of international investigations point to” = “la grande majorité de point d'enquêtes internationales” (84)
10. “i wish the negotiators continued success” = “je voudrais les négociateurs poursuivi avec succès” (91)
11. “opening and closing of chapters” = “l'ouverture et de conclure de chapitres” (96)
12. “that of enlargement” = “que de l'élargissement” (102)
13. “should now be doing their utmost to condemn” = “devraient faire leur plus grande de dénoncer” (108)
14. “to condemn and fight” = “dénoncer et de lutte” (108)
15. “the thoughts and concerns” = “les réflexions et concerne” (109)
16. “brussels 's hold” = “de bruxelles, organiser” (121)

Wrong inflection of lemma

Number of occurrences: 48

Wrong tense

Number of occurrences: 14

Present/verb base form -> infinitive, especially in long sentences

Number of occurrences: 8
Occurrences:

1. “reports of the european court of auditors continually comment” = “[...] continuellement commenter” (14)
2. “we did not, consequently, vote” => “nous n’a pas, par consequent, voter” (15)
3. “we in parliament no longer present” = “nous au parlement non plus présenter” (25)
4. “to see the negotiations proceed” = “que les négociations procéder” (25)
5. “there are too many parties at the moment which [...] display” = “il y a trop de partis à l'heure actuelle qui [...] faire preuve” (79)
6. “the new regulations will merge , harmonise” = “les nouveaux règlements aura merge, harmonise” (124)
7. “new institutional arrangements that allow the community” = “nouveaux accords institutionnels que permettre” (144)

8. “the criteria laid down at the copenhagen council in 1993 on democratisation or the ability to withstand competition actually strengthen” = “les critères [...] renforcer” (149)

Imperative -> Present

Number of occurrences: 1
Occurences:
1. “do not get overwhelmed” = “ne reçoivent pas overwhelmed” (78)

Infinitive -> Present

Number of occurrences: 1
Occurrences:
1. «to acquire land” = “prennent terre” (80)

Gerund (present) -> Gérondif

Number of occurrences: 1
Occurrences
1. “no political group is questioning” = “est interpellant” (85)

Past -> Past participle

Number of occurrences: 3
Occurrences:
1. «the commissioner at long last started” = “le comissaire, à long dispel enfin commencé” (36)
2. “they expressed” = “ils exprimé” (48)
3. “to the bse tragedy that turned into a real saga” = “que l'esb convertie” (136)

Wrong agreements

Ambiguous constructions

Number of occurrences: 1
Occurrences:
1. “independent haulage firms” = “ entreprises de transport routier indépendant” (3)

Subject-participle agreement

Number of occurrences: 18
Occurrences:
1. “that the system cannot be used” = “ le système ne peut être utilisée” (4)
2. “political discourse must always be realistic” = “leur discours politiques doivent toujours être réaliste” (9)
3. “the latter can be concluded” = “ceux-ci peuvent être conclu” (20)

4. “the problems that are raised” = “des problèmes qui sont soulevées” (27)
5. “estonia is concerned” = “l'estonie est concernés” (50)
6. “it has been addressed” = “il a été traitée” (61)
7. “the accession process is judged” = “le processus d'adhésion est juger”
8. “the european union just wants to be sure “ = “l'union européenne seulement veut être certain” (66)
9. “my report is also bound” = “mon rapport est également liée” (74)
10. “the european funds for regional development cannot continue to be granted” = “le fonds européen [...] à être accordées” (74)
11. “it cannot be carried” = “elle ne peut être réalisé” (76)
12. “minorities being given” = “de minorités accordée” (98)
13. “the debate over the order in which the countries are to join the union remains open” = “le débat [...] reste ouverte” (105)
14. “the state of progress of their internal reforms will have to be assessed” = “l'état de progrès [...] doivent être traitée” (105)
15. “rights were properly enshrined” = “les droits ont été correctement contenues” (112)
16. “the fact that such dilution occurs must be made public” = “le fait que [...] doit être rendue publique”(116)
17. “the public also need to be aware” = “le publique doit également être conscients” (116)
18. “the value limit for fish meal , for example , is different” = “la valeur [...] est différent” (120)

Subject-verb agreement

Number of occurrences: 5
Occurrences:

1. «we did not» => «nous n’a pas» (15)
2. “.the report which it is my honour to submit , which was also unanimously adopted in committee , observes” = “le rapport [...] observes” (64)
3. “that the accession process progresses” = “que le processus d'adhésion progresses” (66)
4. “countries who , ten years ago , had” = “les pays qui, il y a dix ans, eu” (68)
5. “compensation is necessary” = “les indemnisations est nécessaire” (127)

Noun – modifier (determiner, adjective)

Number of occurrences: 8
occurrences:

1. «political discourse » = « leur discours politiques» (9)
2. «the real weight” = “la véritable poids” (9)
3. “a quite precise overview” = “un aperçu très précises” (26)
4. “each of them” = “chacune d'entre eux” (27)
5. “the free right” = “la libre droit” (80)
6. “insignificant minorities” = “des minorités insignifiant” (85)
7. “a real saga” = “un véritable histoire” (136)
8. “derived legislation” = “la législation issu” (145)

Reference

Number of occurrences: 2
Occurrences:

1. “, a perfectly legitimate question , but it is one” = “une question tout à fait légitime, mais c'est celui” (28)
2. “the charter of fundamental rights because it summarises” = “la charte des droits fondamentaux parce qu'il summarises” (21)

d) Word not translated (verb, noun, preposition)

Number of occurrences: 9

Occurrences:

1. “the second major change which we are witnessing and to which europe must adjust is the meteoric speed” = “le deuxième grand changement [...] que l'europe doit meteoric ajuster le rythme auquel” (31)
2. “the present members claimed long transitional periods” = “les membres actuels des longues périodes transitoires” (33)
3. “at long last started to dispel” = “à long dispel enfin commencé” (36)
4. “geared more to the citizens” = “s'ouvrir davantage les citoyens” (53)
5. “let us do that” = “faisons” (80)
6. “have insufficient insight” = “on insight dans” (89)
7. “the second form of communication we need is that between the member states” = “l'autre forme de communication que nous avons besoin d'entre les états membres” (92)
8. “have to live up to this challenge” = “à hauteur de ce défi” (95)
9. “the value limit for fish meal” = “la valeur limite farine de poisson” (120)

e) Wrong ordering of modifiers and of agents (subject/object)

Number of occurrences: 27

Noun/adjective or noun compounds

Number of occurrences: 17
Occurrences:

1. “cheaper and internationally deployable” = “au niveau international, déployable moins cher” (16)
2. “general report” = “générale du rapport” (61)
3. “copenhagen political criteria” = “les critères de copenhague politique” (64)
4. “the commission 's own role” = “la commission de la propre rôle” (65)
5. “the current external borders” = “l'actuelle aux frontières extérieures” (75)
6. “heavy freight” = “lourde au transport de marchandise” (76)
7. “too little ambition and too much materialism” = “trop peu et trop d'ambition materialism” (79)
8. “family use” = “famille de l'utilisation” (80)
9. “by a bureaucratic lack of transparency” = “bureaucratique par un manque de transparence” (81)

10. “the institutional , ordered regime” = “ordered institutionnelle, régime” (87)
11. “equal status” = “l'égalité d'état” (99)
12. “a lot of inherited knowledge and farmers ' lore” = “beaucoup de connaissances et des agriculteurs hérité d'lore” (123)
13. “very detailed and complex hygiene requirements” = “très détaillée et complexe d'hygiène” (124)
14. “aliments matériaux” = “feed materials” (127)
15. “scientific risk assessments” = “”risques des évaluations scientifiques” (129)
16. “membership application” = “adhésion à l'application” (142)
17. “membership application” = “adhésion à l'application” (143)

Syntactic reordering

Number of occurrences: 10

1. “the fundamental rights which the public are entitled to” = “les droits fondamentaux qui ont droit à l'opinion publique” (21)
2. “have like to have seen a moratorium” = “comme d'un moratoire sur l'avons vu” (23)
3. “a specific outcome is needed” = “il est indispensable d'un résultat concret” (30)
4. “which does justice to the challenge” = “qui n'est le défi de la justice” (67)
5. “for health impact assessments for all major legislation” = “pour tous les grandes évaluations des incidences sur la législation” (73)
6. “results in many long-term health consequences” = “de nombreux résultats à long terme, les conséquences” (87)
7. “regime which operates in that country results in many long-term health consequences” = “régime [...] dans de nombreux résultats à long terme, les conséquences pour la santé” (87)
8. “many countries have faced similar problems” = “de nombreux pays ont des problèmes identiques face” (97)
9. “there are no grounds for refusing candidate countries” = “il n'y a des pays candidats pour refuser” (145)
10. “at the copenhagen council in 1993 on democratisation” = “à copenhague en 1993, le conseil de démocratisation” (149)

2. French to English Translation

a) Source language words inserted in the translation

Number of occurrences: 83

Occurrences:
traversons, aimablement, reproche, fixons, parte, potentialite, sous-estimations, passager, enumerant, dialogue, rende, differenciee, eludera, repondraient, foudroyante, rattraper, sollicite, longs, reformes, modere, econome, intelligente, montaient, immiscer, anti-nucleaire, reviendra, malte, jugent, fluidite, attachee, prononcons, eurent, descendre, branle, eblouir, pourcentages, transpose, degradant, oncle, situons, fluvial, fluviales, egoisme, populisme, derouter, ideologie, materialisme, differons-le, quoique, insignifiantes, elargissait, desintegrer, prescrit, orphelins, mu, affronte, majorites, dependre, rattrapent, ralliement, combattues, accumulent, dioxine, aflatoxines, mercure, dioxine, imperieuse, fallacieuse, tutelle, justifient,

melanges, eparpillees, dedommagement, parviendront, travaillerait, pourcentages, composes, adoptera, derivee, derives, populistes, fascisantes, pointent

b) Wrong word: translation not understandable without looking at source text

Number of occurrences: 30
Occurrences:

1. " qui tente certains" = "which is certain" (6)
2. "méthodes intergouvernementales" = "methods schemes" (6)
3. " l ' échelle mondiale" = "and the world" (16)
4. "rassemble" = "visiting" (21)
5. "comme le proposent" = "so as to why" (23)
6. "mais se contente de" = "are only of" (25)
7. "qui se présentent" = "which are completely" (27)
8. " commissaire verheugen veut" = " commissioner verheugen is" (44)
9. " sont plus proches" = "are closest" (67)
10. " nous serons tous les bénéficiaires de l ' élargissement" = "we shall all beneficiaries of enlargement" (75)
11. "en faisant remarquer " = "by out" (75)
12. "elle existe jusqu ' à présent" = "it is so far" (81)
13. "font état de gains" = "are state of politics" (84)
14. " du groupe eldr" = "of the group procedure" (86)
15. "je souhaite" = "i hope" (91)
16. "il s'agit" = "it is for" (95)
17. "de manière appliquée" = "applied in" (96)
18. " sur un pied d ' égalité" = "on a more equal" (99)
19. "pour ainsi dire" = "so to" (100)
20. "mais qui augmentent" = "but which are" (122)
21. "savoir-faire ancestral" = "know-how generations" (123)
22. "subissent bel et bien" = "suffer apply" (126)
23. "il devront détruire " = "it will destroy (126)
24. "sur la base de ces différentes évaluations" = "i am leaving the principle" (132)
25. " un véritable roman" = "a real description" (136)
26. " effectivement , efficacement" = "effectively, effectively" (145)
27. " refuser aux pays candidats un droit" = "to reject the applicant countries a right" (145)
28. "le parlement pose un geste politique" = "parliament is a clear political gesture" (147)
29. " le règne d ' un capitalisme" = "it is of a definition wild capitalism" (149)
30. "une fuite en avant" = "on a prior" (150)

c) Wrong word/phrase: translation has something in common with original text

Number of occurrences: 73

Literal translation: Words translated as in source text

Number of occurrences: 31

Expressions and subordinate clauses

Number of occurrences: 12
Occurrences:

1. “ auquel l ' europe doit s ' adapter” = “which europe must adapt” (31)
2. “les états ayant entamé” = “states have initiated” (32)
3. “un groupe à haut niveau” = “a high level group on” (63)
4. “le problème principal auquel malte doit faire face est que [...]” = “the main problem facing malte must face is that [...]” (66)
5. “ exception faite de minorités insignifiantes” = “made insignificant exception” (85)
6. “ne met en question” = “is not in question” (85)
7. “nous avons par conséquent intérêt à étendre” = “we have therefore interest to extend” (101)
8. “ à maintes reprises” = “in repeatedly” (106)
9. “auxquelles elle fait référence” = “which it refers” (122)
10. “du bon sens paysan” = “the right direction farmer” (123)
11. “ que l ' union européenne en tire” = “the european union to draw” (145)
12. “ doit se faire de manière progressive” = “must be gradual manner” (146)

Wrong use of French 'ne… que', 'ne ... plus', 'ne ... pas'

Number of occurrences: 5
Occurrences:

1. “ que le parlement ne présente plus , l ' année prochaine” = “that parliament presents more next year” (25)
2. “la présidence française n ' éludera aucune des difficultés” = “the french presidency is éludera any problems” (27)
3. “n ' avaient pas de raison d ' être” = “had not justification” (36)
4. “ leur adhésion ne doit dépendre que de leur propre développement“ = “membership must not dépendre than their own development”
5. “ les animaux ne sont plus que des unités” = “animals are more than the units” (123)

Impersonal structure: 'Il faut'; 'on'

Number of occurrences: 6
Occurrences:

1. “il faudrait garder à l ' esprit” = “should bear in mind” (33)
2. “ il faut aussi pouvoir recueillir l ' adhésion” = “it must also be able to obtain the support” (45)
3. “ qu ' il me soit également permis” = “it is also enabled me” (68)
4. “on a également parlé” = “it has also referred” (77)
5. “il faudrait la désintégrer” = “it should be the désintégrer” (86)
6. “si l ' on autorise” = “if it permits” (116)

Reflexive pronouns

Number of occurrences: 5

1. “ nous nous attachons” = “we attach” (25)
2. “ afin que nous puissions , en tant qu ' union européenne , nous fixer les objectifs” = “so that we can, as the european union, we have to fulfil the objectives” (39)
3. “ je veux me faire bien comprendre” = “i want to make quite clear” (45)
4. “ c ' est pourquoi nous nous prononçons” = “that is why we prononçons” (67)
5. “ pour me joindre à nombre de confrères” = “to join me in many colleagues” (109)

Same POS but wrong word: wrong preposition

Number of occurrences: 14
Occurrences:

1. “ont à présent une vision” = “have to present a clear vision” (5)
2. “de la grèce au portugal” = “of greece, portugal” (14)
3. “elle est à l ' image de” = “it is for the image of” (22)
4. “nous devons aborder avec beaucoup de générosité” = “we must deal with many with generosity” (29)
5. “dans les domaines où” = “in the fields of considerable investments” (33)
6. “ pour que l ' élargissement” = “to that enlargement” (42)
7. “à notre haut représentant” = “in our high representative” (65)
8. “ pour me joindre à nombre de confrères” = “to join me in many colleagues” (109)
9. “substances à la toxicité violente” = “substances in the toxicity of violent” (114)
10. “ il faut au moins en notifier” = “we must at least in notifier” (116)
11. “pour les farines de poisson” = “for those of fish meal” (120)
12. “ des mesures appropriées” = “of the appropriate measures” (129)
13. “ainsi que mme roth-behrendt” = “as mrs roth-behrendt” (135)
14. “ à plusieurs chapitres” = “in several chapters” (136)

Wrong POS

Number of occurrences: 11

Wrong POS

Number of occurrences: 3
Occurrences:

1. “ cette seule raison suffit à expliquer notre refus” = “the only reason enough to explain” (20)
2. “les préoccupations [...] risquent d ' être exploitées” = “concerns [...] likely to” (62)
3. “ peuvent conseiller la bulgarie” = “can bulgaria adviser” (97)

Noun-adjective wrong POS

Number of occurrences: 4
Occurrences:

1. “ le discours politique” = “the speeches policy” (9)
2. “ son avenir indépendant” = “its future independent” (40)
3. “ énergie politique nécessaire” = “energy policy necessary” (96)
4. “ origine animale non maritime” = “animal origin not sea” (120)

Pronouns/determiners confusion

Number of occurrences: 4
Occurrences :

1. “ m. verheugen l ' ait une fois encore soulignée” = “mr verheugen the has once again” (37)
2. “il faudrait la désintégrer” = “it should be the désintégrer” (86)
3. “ afin de leur expliquer” = “to their explain” (92)
4. “ afin de l ' aider” = “to the help” (97)

Wrong inflection of lemma

Number of occurrences: 15

Wrong tense

Wrong tense: Subjunctive in French or Passive to Active

Number of occurrences: 6
Occurrences:

1. “il faut apprendre , connaître” = “we must learn, knowing” (9)
2. “ il faut absolument respecter” = “we must be respected” (18)
3. “ que le commissaire commence enfin” = “I was happy that the commissioner finally begin” (36)
4. “ j ' ai pu m ' en rendre compte “ = “i am able to realize” (106)
5. “vont regrouper , harmoniser” = “will bring together, harmonizing” (124)
6. “ la législation dérivée que l ' union européenne en tire” = “the legislation dérivée that the european union to draw” (145)

Wrong agreements

Number of occurrences: 9
Occurrences:

1. “ quelles sont les compétences” = “what are the competence” (1)
2. “ ces bus , ces derniers” = “these recent bus” (16)
3. “ des travaux qui ont été engagés” = “business which have been committed” (26)
4. “tous les résultats [...] ne répondraient pas” = “all the results [...] does not” (30)
5. “ je ne joue pas” = “i do not plays” (45)
6. “le rapport [...] constate” = “the report [...] note” (64)
7. “ qui ne sont pas toxiques en soi” = “which are not toxic in itself” (122)
8. “regardons les échanges commerciaux qu ' il y a actuellement” = “look at the trade that currently there are” (35)
9. “la commission a promis qu ' elle” = “the commission had promised that she” (133)

d) Word not translated

Number of occurrences: 4
Occurrences:

1. “ surtout pas nous autres” = “especially not other” (58)
2. “la démocratie slovaque se développe de façon stable” = “democracy is developing in slovak stable” (59)
3. “ sont aussi très soucieux du bien-être” = “are also very concerned the well-being” (127)
4. “de façon correcte et précise” = “a correct and detailed.”

e) Wrong ordering of modifiers and of agents (subject/object)

Number of occurrences: 25

Coordinated NPs including adjectives

Number of occurrences: 11
Occurrences:

1. “ entreprises de transports indépendantes et pour l ' agriculture” = “independent of transport for agriculture” (3)
2. “une charte contraignante et globale” = “a binding charter and comprehensive” (18)
3. “investissements considérables et à long terme” = “considerable investment and long term” (33)
4. “ la sécurité nucléaire et environnementale” = “nuclear safety and environmental” (62)
5. “ pour une stratégie réfléchie , flexible et à la mesure du défi” = “strategy for a considered, flexible and to the extent of the challenge”(67)
6. “ d ' un idéalisme insuffisant et d ' un matérialisme excessif” = “an inadequate and idealism of a matérialisme excessive” (79)
7. “ le manque de transparence bureaucratique et la résistance manifeste” = “lack of transparency and the bureaucratic resistance” (81)
8. “ condamnées et combattues” = “and combattues condemned” (108)
9. “ simplifier des prescriptions sanitaires très détaillées et complexes” = “health requirements and simplifying the very detailed and complex” (124)
10. “ un capitalisme sauvage et destructeur” = “wild capitalism and a destructive” (149)
11. “codécision appropriée” = “codecision appropriate” (98)

Subject/object wrong ordering

Number of occurrences: 3
Occurrences:

1. “la gigantesque pression qu ' exerce l ' élargissement” => “the enourmous pressure that enjoys enlargement” (30)
2. “le deuxième changement majeur [...] est l ' accélération foudroyante” = “the second major change [...] the foudroyante is greater speed” (31)
3. “tout ce que contiennent ces aliments” = “everything contain these foods” (134)

Syntactic wrong reordering on the sentence level

Number of occurrences: 12
Occurrences:

1. “ je répondrai qu ' avant de prononcer un discours” = “that before i respond to give a speech” (9)
2. “ ces derniers sont non seulement meilleur marché et peuvent être utilisés l ' échelle mondiale” = “not only are better market can be used and the world” (16)
3. “le groupe csu au parlement européen se réjouit que” = “the csu group welcomes the european parliament” (21)
4. “ rassemble et rende visibles” = and rende visiting now visible” (21)
5. “je voudrais donc également vous proposer” = “i wish you also propose” (25)
6. “ il découle de cela le principe que les états ayant entamé plus tard” = “it follows that the principle that states have initiated” (32)
7. “ je me réjouis que m. verheugen l ' ait une fois encore soulignée : il ne peut y avoir de délai transitoire” = “i am glad that mr verheugen the has once again: there are highlighted cannot prevail” (37)
8. “ une politique de défense et de sécurité commune européenne , étant donné que cela crée des risques” = “a common defence and security policy, since european this creates” (41)
9. “je me suis vu confier le travail de rapporteur” = “i have seen the work of rapporteur asked” (50)
10. “ jusqu ' en roumanie” = “in romania and until” (101)
11. “il est temps de rompre avec une interprétation fallacieuse du principe de subsidiarité” = “it is time for an interpretation fallacieuse breaking with the principle of subsidiarity” (121)
12. “les critères du conseil de copenhague de 1993 , à propos de la démocratisation ou de la capacité à soutenir la concurrence renforcent” = “the criteria [...] about the ability to strenghten support competition” (149)

B. Legend for Flemm Morphological Analysis

Abbreviations used:
P Position (starts with 0 for encoding PoS values)
ATT Attribute name
VAL Value
C Code

Pronouns (P)

P	ATT	VAL	C
1	Type	personal	p
		demonstrative	d
		indefinite	i
		possessive	s
		interrogative	t
		relative	r
		exclamative	e
		reflexive	x
		reciprocal	l
2	Person	first	1
		second	2
		third	3
3	Gender	masculine	m
		feminine	f
		neuter	n
4	Number	singular	s
		plural	p
5	Case	nominative	n
		genitive	g
		dative	d
		accusative	a
		oblique	o
		object	j
6	Possessor	singular	s
		plural	p

Verbs (V)

P	ATT	VAL	C
1	Type	main	m
		auxiliary	a
		modal	o
2	Mood/VForm	indicative	i
		subjunctive	s
		imperative	m
		conditional	c
		infinitive	n
		participle	p
		gerund	g
		supine	s
		base	b
3	Tense	present	p
		imperfect	i
		future	f
		past	s
4	Person	first	1
		second	2
		third	3
5	Number	singular	s
		plural	p
6	Gender	masculine	m
		feminine	f
		neuter	n

Determiners (D)

P	ATT	VAL	C
1	Type	demonstrative	d
		indefinite	i
		possessive	s
		interrogative	t
2	Person	first	1
		second	2
		third	3
3	Gender	masculine	m
		feminine	f
		neuter	n
4	Number	singular	s
		plural	p
5	Case	nominative	n
		genitive	g
		dative	d
		accusative	a
		oblique	o
6	Possessor	singular	s
		plural	p

Nouns (N)

P	ATT	VAL	C
1	Type	common	c
		proper	p
2	Gender	masculine	m
		feminine	f
		neuter	n
3	Number	singular	s
		plural	p
4	Case	nominative	n
		genitive	g
		dative	d
		accusative	a

Adjectives (A)

P	ATT	VAL	C
1	Type	qualificative	f
		ordinal	o
		cardinal	c
		indefinite	i
		possessive	s
2	Degree	positive	p
		comparative	c
		superlative	s
3	Gender	masculine	m
		feminine	f
		neuter	n
4	Number	singular	s
		plural	p
5	Case	nominative	n
		genitive	g
		dative	d
		accusative	a

Articles (T)

P	ATT	VAL	C
1	Type	definite	d
		indefinite	i
2	Gender	masculine	m
		feminine	f
		neuter	n
3	Number	singular	s
		plural	p
4	Case	nominative	n
		genitive	g
		dative	d
		accusative	a

Adverbs (R)

P	ATT	VAL	C
1	Type	general	g
		particle	p
2	Degree	positive	p
		comparative	c
		superlative	s

Appositions (S)

P	ATT	VAL	C
1	Type	preposition	p
		postposition	t
		circumposition	c
2	Formation	simple	s
		compound	c

Conjunctions (C)

P	ATT	VAL	C
1	Type	coordinating	c
		subordinating	s

Numerals (M)

P	ATT	VAL	C
1	Type	cardinal	c
		ordinal	o
2	Gender	masculine	m
		feminine	f
		neuter	n
3	Number	singular	s
		plural	p
5	Case	nominative	n
		genitive	g
		dative	d
		accusative	a

C. Rules for Morphological Disambiguation of French

```
for tags in analysed.txt:
        search ambiguous pronoun
        search ambiguous verb
        search ambiguous noun
I       if ambiguous pronoun:
                search 'Pp\d-p--'           ## nous, vous
                replace lemma by 'il'
                search 'Pp3mpj-'             ## eux
                replace morph by 'Pp3mpj-'
                replace lemma by 'lui'
                search 'Pp3fsj-'            ## elle (lui/elle)
```

```
            replace morph by 'Pp3fsn-'
            search 'Pp3fpj-'            ## elles
            replace morph by 'Pp3fpn-'
            search 'Pp\d-s-'             ## moi, toi
            replace lemma by 'lui'
            search 'Pp1-sj-'             ## me
            replace lemma by 'se'
      if ambiguous noun:
            pick first analysis
      if ambiguous verb:
             search 'Vmip1s--1'              ## signifie
            replace morph by 'Vmip1/3s--1'
            search 'Vmip1s--2'              ## applaudis
             replace morph by 'Vmip/s1/2s--2'
            search 'Vmip1p--1'              ## allons
            replace morph by 'Vmi/mp1p--1'
            search 'Vmip2p--1'              ## trouvez
            replace morph by 'Vmi/mp2p--1'
            search 'Vmip3p--1'              ## incitent
            replace morph by 'Vmip3p--1'
            search 'Vmip3s--2'              ## agit
            replace morph by 'Vmip/s3s--2'
            search 'Vmip1p--2'              ## subissons
            replace morph by 'Vmi/mp1p--2'
            search 'Vmip2p--2'              ## agissez
            replace morph by 'Vmi/mp2p--2'
            search 'V[a|m]ip1s--3_suivre'        ## être
            replace morph by 'Vmip1s--3'
            search '[C|c]rois_VER\(pres\):Vmip1s--3'   ## crois
            replace morph by 'Vmip1s—3'
      replace lemma by 'croire'
            search 'Vmip1s--3'              ## dois, interviens
            replace morph by 'Vmip1s--3'
            search  'Vmip1p--3'                        ##  pouvons,  disons,  faisons,
entendons
            replace morph by 'Vmi/mp1p--3'
            search 'Vmip2p--3'              ## savez, permettez
            replace morph by 'Vmi/mp2p--3'
            search '[S|s]oit_VER\((aux:)subp\):V[a|m]ip3s'  ## wrong tag: soit
            replace morph by 'Vmis3s--3'
            search 'Vmip3s--3'              ## convient
            replace morph by 'Vmip3s--3'
            search 'Vmip1s--3'              ## souscris, fais, comprends
            replace morph by 'Vmip1/2s--3'
            search 'Vmmp2p--3'               ## soyez
            replace morph by 'Vmm/sp2p--3'
            search 'Vmsp1s--3'              ## fasse
            replace morph by 'Vmsp1/3s--3'
            search 'Vmsp1s--2'              ## jouisse
            replace morph by 'Vmsp1/3s--2'
```

```
search 'Vmcp1s--3'                  ## voudrais
replace morph by 'Vmcp1/2s--3'
search 'Vmii1p--1'                  ## parlions
replace morph by 'Vmi/si/p1p--1'
search 'Vmii1p--3'                 ## soutenions
replace morph by 'Vmi/si/p1p--3'
search  'Vmii1p--2'                 ## reunissions
replace morph by 'Vmi/si/p1p--2'
search 'Vmii2p--1'                  ## parliez
replace morph by 'Vmi/si/p2p--1'
search 'Vmii2p--3'                  ## souteniez
replace morph by 'Vmi/si/p2p--3'
search 'Vmii2p--2'                  ## fournissiez
replace morph by 'Vmi/si/p2p--2'
search 'Vaii3p--3'              ## étaient
pick first analysis
search 'Vmip3p--3'                  ## doivent
replace morph by 'Vmi/sp3p--3'
search 'Vmip3p--2'                 ## unissent
replace morph by 'Vmi/sp3p--2'
search 'Vmcp1s--1'                  ## souhaiterais
replace morph by 'Vmcp1/2s--1'
search 'Vmcp1s--2'                  ## saisirais
replace morph by 'Vmcp1/2s--2'
search 'Vacp1s--3'                 ## serais
replace morph by 'Vacp1/2s--3'
search 'Vamp1p--3'                  ## ayons
replace morph by 'Vam/sp1p--3'
search '[E|e]tes_VER\(pres\):Vmip2s--1'      ## êtes
replace morph by 'Vaip2p--3'
replace lemma by 'être'
search 'Vmmp1p--3'                  ## soyons, ayons
replace morph by 'Vmm/sp1p--3'
search 'Vamp2p--3'                  ## soyez
replace morph by 'Vam/sp2p--3'
search 'Vmii1s--3'                 ## avais (main verb)
replace morph by 'Vmii1/2s--3'
search 'Vaii1s--3'                ## avais (aux)
replace morph by 'Vaii1/2s--3'
search 'Vmii1s--1'                ## trouvais
replace morph by 'Vmii1/2s--1'
search 'Vmii1s--2'                ## agissais
replace morph by 'Vmii1/2s--2'
search 'Vacp1s--3'                  ## aurais
replace morph by 'Vacp1/2s--3'
search 'Vamp2s--3'                  ## aie
replace morph by 'Vasp1s--3'
```

D.Manual Evaluation –Error Distributions of Experiments

1. Distribution of errors – Combined Feature Models Error Report: English to French Translations

a) Word and lemma

	Total occurrences	Percentage out of total words in output
English words inserted in translation	**96**	**2.29%**
Wrong words translation	**33**	**0.78%**

	Occurrences	Percentage out of total occurrences of same expression type
Literal translations of fixed expressions	60	54% (out of total fixed expressions)
Same POS but wrong word	28	5.69% (of total prepositions)
Wrong POS	20	0.47% (of total nb of words)
Wrong tense	26	8.1% (of verb phrases)
Wrong agreements •Noun-modifier •Subject-Verb /Participle	33 10 19	2.68% (of verb agreement)
Missing words	7	0.16% (out of total words in output)
Total	**203**	

	Occurrences	Percentage out of total occurrences of same expression type
Noun Phrases including adjectives	26	9.6% (of all NPs including adjectives)
Wrong ordering at the sentence level	8	2.5% (of total nb of clauses)
Total	**34**	

b) Word and POS

	Total occurrences	Percentage out of total words in output
English words inserted in translation	**140**	**3.32%**
Wrong words translation	**27**	**0.64%**

	Occurrences	Percentage out of total occurrences of same expression type
Literal translations of fixed expressions	48	43.23% (out of total fixed expressions)
Same POS but wrong word	32	6.5% (of total prepositions)
Wrong POS	15	0.35% (of total nb of words)
Wrong tense	13	4% (of verb phrases)
Wrong agreements • Noun-modifier • Subject-Verb /Participle	49 9 35	3.98% (of verb agreement)
Missing words	10	0.23% (out of total words in output)
Total	**167**	

	Occurrences	Percentage out of total occurrences of same expression type
Noun Phrases including adjectives	24	8.92% (of all NPs including adjectives)
Wrong ordering at the sentence level	5	1.56% (of total nb of clauses)
Total	**29**	

c) Word and morphology

	Total occurrences	Percentage out of total words in output
English words inserted in translation	**139**	**3.3%**
Wrong words translation	**40**	**0.9%**

	Occurrences	Percentage out of total occurrences of same expression type
Literal translations of fixed expressions	56	50.4% (out of total fixed expressions)
Same POS but wrong word	38	7.72% (of total prepositions)
Wrong POS	16	0.38% (of total nb of words)
Wrong tense	14	4.3% (of verb phrases)
Wrong agreements • Noun-modifier • Subject-Verb /Participle	35 10 20	2.84% (of verb agreement)
Missing words	15	0.35% (out of total words in output)
Total	**174**	

	Occurrences	Percentage out of total occurrences of same expression type
Noun Phrases including adjectives	26	9.6% (of all NPs including adjectives)
Wrong ordering at the sentence level	3	0.9% (of total nb of clauses)
Total	**29**	

2. Distribution of errors – Factored Models Error Report: French to English Translations

a) Word to word translation, word to POS generation

	Total occurrences	Percentage out of total words in output
French words inserted in translation	**81**	**2.08%**
Wrong words translation	**27**	**0.69%**

	Occurrences	Percentage out of total occurrences of same expression type
Literal translations of fixed expressions	35	16.74% (out of total fixed expressions)
Same POS but wrong word	14	2% (of total prepositions)
Wrong POS	9	0.2% (of total nb of words)
Wrong tense	10	2.99% (of verb phrases)
Wrong agreements	6	1.79% (of verb agreement)
Missing words	22	0.56% (out of total words in output)
Total	**96**	

	Occurrences	Percentage out of total occurrences of same expression type
Noun Phrases including adjectives	25	13.81% (of all NPs including adjectives)
Wrong ordering at the sentence level	9	2.68% (of total nb of clauses)
Total	**30**	

b) Word to word translation, word to POS and lemma generation

	Total occurrences	Percentage out of total words in output
French words inserted in translation	**79**	**2%**
Wrong words translation	**23**	**0.59%**

	Occurrences	Percentage out of total occurrences of same expression type
Literal translations of fixed expressions	32	15.3% (out of total fixed expressions)
Same POS but wrong word	12	1.79% (of total prepositions)
Wrong POS	11	0.28% (of total nb of words)
Wrong tense	5	1.49% (of verb phrases)
Wrong agreements	7	2% (of verb agreement)
Missing words	20	0.51% (out of total words in output)
Total	**87**	

	Occurrences	Percentage out of total occurrences of same expression type
Noun Phrases including adjectives	21	11.6% (of all NPs including adjectives)
Wrong ordering at the sentence level	8	2.39% (of total nb of clauses)
Total	**29**	

E. Unseen words inserted in output translation tables

1. Combined feature models

		Training size	Occurrences (%)	Number of word types	Types with morpho variant in training data
French to English Translations	Baseline (Words)	Limited corpus	6.1	1923	1157
		Whole corpus	3.9	851	442
	Word + POS	Limited corpus	6.6	2110	1287
		Whole corpus	4	901	480
	Word + Morph	Limited corpus	6.5	2047	1241
		Whole corpus	4	898	477
	Word + POS + Morph	Limited corpus	6.5	2088	1261
		Whole corpus	4	914	494
	Word + Lemma	Limited corpus	6.2	1952	1175
		Whole corpus	3.9	857	445
English to French Translations	Baseline (Words)	Limited corpus	4.7	1656	1082
		Whole corpus	3.2	806	682
	Word + POS	Limited corpus	6.1	2067	1393
		Whole corpus	4.2	1074	914
	Word + Morph	Limited corpus	6.1	2061	1390
		Whole corpus	4.2	1074	914
	Word + POS Morph	Limited corpus	6.1	2071	1399
		Whole corpus	4.2	1080	917
	Word + Lemma	Limited corpus	5.2	1726	1133
		Whole corpus	3.5	843	709

2. Factored models

		Training size	Occurrences (%)	Number of word types	Number of types with morpho variant in training data
French to English Translations	Baseline (Words)	Limited corpus	6	1905	1142
		Whole corpus	3.9	847	438
	Word → POS	Limited corpus	6	1905	1142
		Whole corpus	3.9	854	443
	Word → Lemma	Limited corpus	6.1	1907	1143
		Whole corpus	3.9	855	444
	Word → POS, Lemma	Limited corpus	6.1	1908	1146
		Whole corpus	3.9	862	450
English to French Translations	Baseline (Words)	Limited corpus	4.8	1667	1094
		Whole corpus	3.3	818	690
	Word → POS	Limited corpus	4.9	1655	1086
		Whole corpus	3.3	807	677
	Word → Morph	Limited corpus	4.8	1665	1091
		Whole corpus	3.4	825	694
	Word → POS, Morph	Limited corpus	4.9	1661	1089
		Whole corpus	3.4	823	690
	Word → Lemma	Limited corpus	4.9	1656	1083
		Whole corpus	3.3	821	693
	Word → Morph, Lemma	Limited corpus	4.9	1687	1109
		Whole corpus	3.3	801	676
	Word → POS, Lemma	Limited corpus	4.9	1662	1087
		Whole corpus	3.4	822	697

F. Mistranslation errors by POS

1. Mistranslation Summary For Combined Features

In the following tables, the numbers are percentages of total words for the same POS category.

Error report for Pharaoh baseline, English to French translations, trained on limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 50.2476780186 NOM: 34.8188803513 PUN: 100.0 VER: 53.3349444176 KON: 21.3585179804 INT: 63.6363636364 PRO: 37.9002079002 DET: 15.0841750842 ABR: 50.0 NUM: 26.5 PRP: 39.0394877268 PRE: 63.6363636364 ADJ: 39.5219308424	ADV: 2.72445820433 NOM: 2.70032930845 VER: 16.4088931851 KON: 3.99564111878 PRO: 16.237006237 DET: 36.8518518519 NUM: 0.5 PRP: 7.79082177161 ADJ: 11.6268275702

Error report for Pharaoh baseline, English to French translations, trained on the whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 48.6996904025 NOM: 30.6988657153 PUN: 44.8818897638 VER: 49.2266795553 KON: 21.5038140211 INT: 45.4545454545 PRO: 36.237006237 DET: 14.4949494949 ABR: 50.0 NUM: 25.0 PRP: 35.8164354322 PRE: 63.6363636364 ADJ: 34.9268971919	ADV: 2.16718266254 NOM: 2.58324185876 VER: 16.0463992267 KON: 3.77769705776 PRO: 16.1330561331 DET: 35.8585858586 NUM: 0.5 PRP: 7.69477054429 ADJ: 11.093061035

Error report for Combined Word and Lemma Models, English to French translations, trained on the limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 51.6099071207 NOM: 35.0896450787 PUN: 76.3779527559 VER: 53.4195263412 KON: 22.339266255 INT: 54.5454545455 PRO: 38.5446985447 DET: 15.0 ABR: 100.0 NUM: 25.5 PRP: 38.1216648879 PRE: 81.8181818182 ADJ: 38.6400556974	ADV: 2.16718266254 NOM: 2.634467618 VER: 16.2155630739 KON: 2.57900472212 PRO: 15.343035343 DET: 35.7070707071 PRP: 6.13660618997 ADJ: 12.0909723834

Error report for Combined Word and Lemma Models, English to French translations, trained on the whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 50.3095975232 NOM: 31.2916209294 PUN: 36.2204724409 VER: 49.6858385694 KON: 21.8307301126 INT: 45.4545454545 PRO: 36.6528066528 DET: 13.9057239057 ABR: 100.0 NUM: 24.5 PRP: 34.7385272145 PRE: 72.7272727273 ADJ: 35.1821768392	ADV: 2.16718266254 NOM: 2.53933406513 VER: 15.5389076849 KON: 2.68797675263 PRO: 15.1351351351 DET: 35.6565656566 NUM: 0.5 PRP: 6.49946638207 ADJ: 11.1394755164

Error report for Combined Word and Morphology Models, English to French translations, trained on the limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 52.3529411765 NOM: 37.0215879985 PUN: 76.3779527559 VER: 54.6761720638 KON: 22.8478023974 INT: 54.5454545455 PRO: 37.920997921 DET: 15.303030303 ABR: 50.0 NUM: 30.0 PRP: 38.4738527215 PRE: 81.8181818182 ADJ: 43.4671617545	ADV: 2.13622291022 NOM: 2.66373948042 VER: 15.3455775737 KON: 2.28841264076 PRO: 15.0935550936 DET: 36.3468013468 NUM: 0.5 PRP: 5.92315901814 ADJ: 11.4875841262

Error report for Combined Word and Morphology Models, English to French translations, trained on the whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 49.1950464396 NOM: 31.8331503842 PUN: 35.4330708661 VER: 50.1208313195 KON: 21.9033781329 INT: 63.6363636364 PRO: 37.3596673597 DET: 14.3602693603 ABR: 100.0 NUM: 26.5 PRP: 35.7844183565 PRE: 72.7272727273 ADJ: 35.9712230216	ADV: 2.29102167183 NOM: 2.53201609952 VER: 15.8530691155 KON: 2.32473665093 PRO: 14.3866943867 DET: 35.4882154882 PRP: 5.86979722519 ADJ: 11.371547923

Error report for Combined Word and POS Models, English to French translations, trained on the limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 51.8575851393 NOM: 36.7874130992 PUN: 76.3779527559 VER: 54.4224262929 KON: 22.7388303669 INT: 54.5454545455 PRO: 37.6299376299 DET: 15.5387205387 ABR: 100.0 NUM: 29.0 PRP: 38.8580576307 PRE: 81.8181818182 ADJ: 42.3300069622	ADV: 2.41486068111 NOM: 2.61251372119 VER: 15.3939101015 KON: 2.76062477297 PRO: 14.9688149688 DET: 35.7407407407 NUM: 0.5 PRP: 5.94450373533 ADJ: 12.1373868647

Error report for Combined Word and POS Models, English to French translations, trained on the whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 49.2879256966 NOM: 31.7965605562 PUN: 37.7952755906 VER: 50.1812469792 KON: 22.1939702143 INT: 63.6363636364 PRO: 37.525987526 DET: 14.1414141414 ABR: 100.0 NUM: 26.5 PRP: 35.7737459979 PRE: 72.7272727273 ADJ: 36.1568809469	ADV: 2.29102167183 NOM: 2.54665203074 VER: 15.7201546641 KON: 2.43370868144 PRO: 14.5945945946 DET: 35.3703703704 PRP: 6.00853788687 ADJ: 11.371547923

Error report for Combined Word, POS and Morphology Models, English to French translations, trained on the limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 51.2383900929 NOM: 36.7215514087 PUN: 76.3779527559 VER: 54.5190913485 KON: 22.7025063567 INT: 54.5454545455 PRO: 38.0249480249 DET: 15.8922558923 ABR: 50.0 NUM: 29.5 PRP: 39.1889007471 PRE: 81.8181818182 ADJ: 41.4249245765	ADV: 2.22910216718 NOM: 2.61251372119 VER: 15.6114064766 KON: 2.50635670178 PRO: 14.7609147609 DET: 35.5555555556 NUM: 0.5 PRP: 5.80576307364 ADJ: 11.9053144581

Error report for Combined Word, POS and Morphology Models, English to French translations, trained on the whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 49.0092879257 NOM: 31.9282839371 PUN: 36.2204724409 VER: 50.0724987917 KON: 22.339266255 INT: 63.6363636364 PRO: 37.2765072765 DET: 14.4107744108 ABR: 100.0 NUM: 27.0 PRP: 36.0085378869 PRE: 72.7272727273 ADJ: 35.739150615	ADV: 2.22910216718 NOM: 2.59787778997 VER: 15.9859835669 KON: 2.32473665093 PRO: 14.5322245322 DET: 35.5723905724 PRP: 5.92315901814 ADJ: 11.4179624043

Error report for Pharaoh baseline, French to English translations, trained on the limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 54.7309833024 VB: 60.24428684 ": 100.0 RB: 54.068914956 NN: 38.5198135198 FW: 68.4210526316 CC: 17.7715877437 WP: 49.1428571429 CD: 32.8431372549 TO: 37.8422782037 NNPS: 12.8342245989 PRP: 34.7252747253 JJ: 46.0851823514 IN: 40.2204597392 NNP: 14.806312769 DT: 24.8797056326 VBinflec: 71.872808043 UH: 50.0 EX: 59.8130841121	MD: 0 VB: 8.39243498818 NNPS: 2.67379679144 RB: 0 NN: 3.37024087024 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.510073960724 IN: 0 DT: 0 VBinflec: 17.6057984569 NNP: 0.889526542324

Error report for Pharaoh baseline, French to English translations, trained on the whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 52.5974025974 VB: 54.3341213554 ": 100.0 RB: 51.4296187683 NN: 34.6736596737 FW: 47.3684210526 CC: 18.3844011142 WP: 46.5714285714 CD: 33.3333333333 TO: 35.2683461117 NNPS: 10.6951871658 PRP: 33.36996337 JJ: 41.8515684774 IN: 38.6611103643 NNP: 10.9899569584 DT: 24.6674214549 VBinflec: 66.0509703063 UH: 50.0 EX: 55.1401869159	MD: 0 VB: 8.70764381403 NNPS: 4.27807486631 RB: 0 NN: 3.43822843823 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.408059168579 IN: 0 DT: 0 VBinflec: 18.0500350713 NNP: 0.659971305595

Error report for Combined Word and Lemma models, French to English translations, trained on the limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 55.4730983302 VB: 59.8502758077 ": 100.0 RB: 54.215542522 NN: 38.7626262626 FW: 73.6842105263 CC: 18.1615598886 WP: 48.2857142857 CD: 38.2352941176 TO: 40.4162102957 NNPS: 13.9037433155 PRP: 33.956043956 JJ: 45.9576638613 IN: 40.9463637586 NNP: 15.0932568149 DT: 24.4126804416 VBinflec: 72.1767594108 UH: 50.0 EX: 62.6168224299	MD: 0 VB: 8.6682427108 NNPS: 3.74331550802 RB: 0 NN: 3.40909090909 CC: 0 WP: 0 CD: 0.490196078431 TO: 0 PRP: 0 JJ: 0.408059168579 IN: 0 DT: 0 VBinflec: 17.6759410802 NNP: 0.918220946915

Error report for Combined Word and Lemma models, French to English translations, trained on the whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 51.4842300557 VB: 53.5460992908 ": 100.0 RB: 53.0058651026 NN: 34.5182595183 FW: 47.3684210526 CC: 18.1615598886 WP: 46.8571428571 CD: 30.8823529412 TO: 35.3231106243 NNPS: 10.6951871658 PRP: 32.673992674 JJ: 41.9790869676 IN: 38.9299637048 NNP: 10.7030129125 DT: 23.6060005661 VBinflec: 65.9574468085 UH: 50.0 EX: 55.1401869159	MD: 0 VB: 8.6682427108 NNPS: 4.8128342246 RB: 0 NN: 3.48679098679 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.382555470543 IN: 0 DT: 0 VBinflec: 18.1201776946 NNP: 0.688665710187

Error report for Combined Word and Morphology models, French to English translations, trained on the limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 55.2875695733 VB: 60.1260835303 ": 100.0 RB: 54.435483871 NN: 38.8014763015 FW: 84.2105263158 CC: 18.217270195 WP: 49.4285714286 CD: 37.2549019608 TO: 39.8685651698 NNPS: 13.3689839572 PRP: 33.0036630037 JJ: 46.2382045397 IN: 41.7663664471 NNP: 15.7532281205 DT: 24.341919049 VBinflec: 72.4339490297 UH: 33.3333333333 EX: 61.6822429907	MD: 0 VB: 7.84081954295 NNPS: 3.20855614973 RB: 0 NN: 3.33139083139 CC: 0 WP: 0 CD: 0.490196078431 TO: 0 PRP: 0 JJ: 0.561081356797 IN: 0 DT: 0 VBinflec: 17.7460837035 NNP: 0.918220946915

Error report for Combined Word and Morphology models, French to English translations, trained on the whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 50.0927643785 VB: 53.9007092199 ": 100.0 RB: 52.6392961877 NN: 34.5571095571 FW: 47.3684210526 CC: 18.1615598886 WP: 46.8571428571 CD: 31.3725490196 TO: 35.2683461117 NNPS: 11.7647058824 PRP: 33.0036630037 JJ: 41.2649834226 IN: 39.024062374 NNP: 10.7890961263 DT: 23.4644777809 VBinflec: 66.6354921674 UH: 50.0 EX: 55.1401869159	MD: 0 VB: 8.47123719464 NNPS: 4.27807486631 RB: 0 NN: 3.29254079254 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.331548074471 IN: 0 DT: 0 VBinflec: 18.3306055646 NNP: 0.631276901004

Error report for Combined Word and POS models, French to English translations, trained on the limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 54.7309833024 VB: 60.24428684 ": 100.0 RB: 54.288856305 NN: 38.6849261849 FW: 84.2105263158 CC: 18.1615598886 WP: 48.8571428571 CD: 38.7254901961 TO: 38.8280394304 NNPS: 14.9732620321 PRP: 33.5164835165 JJ: 46.467737822 IN: 41.4168571044 NNP: 16.0114777618 DT: 24.4551372771 VBinflec: 72.5508534019 UH: 50.0 EX: 62.6168224299	MD: 0 VB: 7.95902285264 NNPS: 3.20855614973 RB: 0 NN: 3.42851592852 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.586585054833 IN: 0 DT: 0 VBinflec: 17.1849427169 NNP: 0.946915351506

Error report for Combined Word and POS models, French to English translations, trained on the whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 50.0 VB: 54.2947202522 ": 100.0 RB: 52.2360703812 NN: 34.8193473193 FW: 47.3684210526 CC: 18.1615598886 WP: 46.5714285714 CD: 31.3725490196 TO: 34.3373493976 NNPS: 10.6951871658 PRP: 32.967032967 JJ: 41.5965314971 IN: 39.0778330421 NNP: 10.9612625538 DT: 23.6060005661 VBinflec: 66.3315407996 UH: 50.0 EX: 57.9439252336	MD: 0 VB: 8.86524822695 NNPS: 5.34759358289 RB: 0 NN: 3.28282828283 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.331548074471 IN: 0 DT: 0 VBinflec: 18.821603928 NNP: 0.602582496413

Error report for Combined Word, POS and Morphology models, French to English translations, trained on the limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 55.5658627087 VB: 59.9290780142 ": 100.0 RB: 54.5454545455 NN: 39.0345765346 FW: 84.2105263158 CC: 18.1058495822 WP: 50.5714285714 CD: 34.3137254902 TO: 37.0755750274 NNPS: 13.9037433155 PRP: 33.2967032967 JJ: 46.0341749554 IN: 41.2958731012 NNP: 15.4662840746 DT: 23.5776960091 VBinflec: 73.2288987608 UH: 50.0 EX: 61.6822429907	MD: 0 VB: 8.98345153664 NNPS: 4.27807486631 RB: 0 NN: 3.37024087024 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.510073960724 IN: 0 DT: 0 VBinflec: 17.3953705869 NNP: 0.832137733142

Error report for Combined Word, POS and Morphology models, French to English translations, trained on the whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 49.814471243 VB: 53.7037037037 ": 100.0 RB: 52.0894428152 NN: 34.7416472416 FW: 47.3684210526 CC: 18.0501392758 WP: 47.1428571429 CD: 31.3725490196 TO: 33.9539978094 NNPS: 10.6951871658 PRP: 32.7106227106 JJ: 41.2904871206 IN: 39.0106197069 NNP: 10.9612625538 DT: 23.6060005661 VBinflec: 66.5653495441 UH: 50.0 EX: 57.0093457944	MD: 0 VB: 8.78644602049 NNPS: 5.34759358289 RB: 0 NN: 3.3411033411 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.331548074471 IN: 0 DT: 0 VBinflec: 18.4475099369 NNP: 0.602582496413

2. Mistranslation Summary for Factored Models

In the following tables, the numbers are percentages of total words for the same POS category.

Error report for Moses baseline, English to French translations, trained on limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
NOM: 36.3190633004 PUN: 96.062992126 VER: 53.8424359594 KON: 22.4119142753 INT: 63.6363636364 PRO: 36.8191268191 DET: 16.0606060606 ABR: 50.0 NUM: 23.0 PRP: 37.8655282818 PRE: 63.6363636364 ADJ: 38.0134601996	ADV: 0.216718266254 NOM: 2.97841200146 VER: 17.3755437409 PRO: 13.3887733888 DET: 29.8316498316 NUM: 0.5 PRP: 2.83884738527 ADJ: 13.274541657

Error report for Moses baseline, English to French translations, trained on whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 48.3281733746 NOM: 30.742773509 PUN: 35.4330708661 VER: 51.5708071532 KON: 22.4119142753 INT: 54.5454545455 PRO: 37.6923076923 DET: 14.4781144781 ABR: 100.0 NUM: 24.0 PRP: 37.4172892209 PRE: 63.6363636364 ADJ: 36.1336737062	ADV: 1.2693498452 NOM: 2.7588730333 VER: 16.8076365394 KON: 0.871776244097 PRO: 14.1372141372 DET: 35.8922558923 PRP: 4.5144076841 ADJ: 13.204919935

Error report for Factored word to lemma model, English to French translations, trained on limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 49.4736842105 NOM: 35.7116721551 PUN: 100.0 VER: 54.8815853069 KON: 22.4119142753 INT: 63.6363636364 PRO: 38.0665280665 DET: 15.5555555556 ABR: 100.0 NUM: 26.5 PRP: 42.5933831377 PRE: 63.6363636364 ADJ: 41.3553028545	ADV: 1.60990712074 NOM: 2.68569337724 VER: 17.5326244563 KON: 0.472212132219 PRO: 14.4698544699 DET: 37.9797979798 PRP: 3.38313767343 ADJ: 16.9644929218

Error report for Factored word to lemma model, English to French translations, trained on whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 48.9783281734 NOM: 30.8232711306 PUN: 33.8582677165 VER: 51.0995650072 KON: 22.4482382855 INT: 36.3636363636 PRO: 36.6112266112 DET: 14.1750841751 ABR: 50.0 NUM: 23.5 PRP: 36.9370330843 PRE: 72.7272727273 ADJ: 35.9016012996	ADV: 1.20743034056 NOM: 2.75155506769 VER: 16.9284678589 KON: 0.762804213585 PRO: 14.4698544699 DET: 35.9764309764 PRP: 4.95197438634 ADJ: 12.9496402878

Error report for Factored word to morphology model, English to French translations, trained on limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 48.6068111455 NOM: 37.2264910355 PUN: 100.0 VER: 54.3740937651 KON: 22.7388303669 INT: 63.6363636364 PRO: 38.3575883576 DET: 15.3367003367 ABR: 100.0 NUM: 25.5 PRP: 38.0789754536 PRE: 63.6363636364 ADJ: 40.7983290787	ADV: 0.123839009288 NOM: 3.00036589828 VER: 16.8197196713 PRO: 13.5135135135 DET: 30.4545454545 PRP: 2.59338313767 ADJ: 12.5783244372

Error report for Factored word to morphology model, English to French translations, trained on whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 48.8854489164 NOM: 30.7061836809 PUN: 36.2204724409 VER: 51.4378927018 KON: 22.4482382855 INT: 36.3636363636 PRO: 37.7754677755 DET: 14.5791245791 ABR: 100.0 NUM: 24.5 PRP: 38.7833511206 PRE: 63.6363636364 ADJ: 36.5978185194	ADV: 1.60990712074 NOM: 2.634467618 VER: 16.7109714838 KON: 1.45296040683 PRO: 14.2827442827 DET: 37.4915824916 PRP: 5.7310565635 ADJ: 12.6943606405

Error report for Factored word to morphology and lemma model, English to French translations, trained on limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 47.5232198142 NOM: 35.8214416392 PUN: 100.0 VER: 53.3470275495 KON: 22.3029422448 INT: 63.6363636364 PRO: 37.5883575884 DET: 14.4612794613 ABR: 100.0 NUM: 23.5 PRP: 37.8014941302 PRE: 63.6363636364 ADJ: 39.0809932699	ADV: 0.0928792569659 NOM: 2.83937065496 VER: 17.2305461576 PRO: 13.5966735967 DET: 30.7744107744 NUM: 0.5 PRP: 2.33724653148 ADJ: 12.9960547691

Error report for Factored word to morphology and lemma model, English to French translations, trained on whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 48.9783281734 NOM: 31.4672521039 PUN: 37.0078740157 VER: 53.5161913968 KON: 21.6491100618 INT: 45.4545454545 PRO: 38.6070686071 DET: 13.7037037037 ABR: 100.0 NUM: 24.5 PRP: 36.2753468517 PRE: 63.6363636364 ADJ: 38.5008122534	ADV: 0.990712074303 NOM: 2.7588730333 VER: 15.7684871919 KON: 0.653832183073 PRO: 13.5135135135 DET: 36.9696969697 PRP: 4.41835645678 ADJ: 11.2555117197

Error report for Factored word to POS model, English to French translations, trained on limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 51.3931888545 NOM: 37.5484815221 PUN: 100.0 VER: 54.3378443693 KON: 23.3563385398 INT: 63.6363636364 PRO: 37.4428274428 DET: 14.6632996633 ABR: 100.0 NUM: 24.0 PRP: 38.8900747065 PRE: 63.6363636364 ADJ: 39.4058946391	ADV: 0.15479876161 NOM: 2.81009879254 VER: 16.8559690672 PRO: 13.4719334719 DET: 30.9259259259 PRP: 2.47598719317 ADJ: 13.5066140636

Error report for Factored word to lemma model, English to French translations, trained on whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 49.4117647059 NOM: 30.9769484083 PUN: 38.5826771654 VER: 51.8003866602 KON: 22.9930984381 INT: 45.4545454545 PRO: 37.5675675676 DET: 13.7878787879 ABR: 100.0 NUM: 23.0 PRP: 37.6093916756 PRE: 63.6363636364 ADJ: 36.6674402414	ADV: 1.5479876161 NOM: 2.73691913648 VER: 16.6868052199 KON: 1.63458045768 PRO: 14.2411642412 DET: 37.138047138 PRP: 5.30416221985 ADJ: 12.8336040845

Error report for Factored word to POS and lemma model, English to French translations, trained on limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 50.7430340557 NOM: 36.2092938163 PUN: 100.0 VER: 52.9482841953 KON: 22.339266255 INT: 63.6363636364 PRO: 36.0083160083 DET: 15.4377104377 ABR: 100.0 NUM: 23.5 PRP: 38.9007470651 PRE: 63.6363636364 ADJ: 38.3151543282	ADV: 0.185758513932 NOM: 2.81741675814 VER: 18.0038666022 PRO: 13.4511434511 DET: 30.6565656566 PRP: 2.47598719317 ADJ: 13.6690647482

Error report for Factored word to POS and lemma model, English to French translations, trained on whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 48.7306501548 NOM: 31.6867910721 PUN: 40.157480315 VER: 49.6858385694 KON: 21.7944061024 INT: 63.6363636364 PRO: 35.1975051975 DET: 13.0976430976 ABR: 50.0 NUM: 25.5 PRP: 34.8986125934 PRE: 63.6363636364 ADJ: 33.3488048271	ADV: 0.15479876161 NOM: 2.97109403586 VER: 17.4963750604 PRO: 13.5550935551 DET: 30.2188552189 PRP: 2.61472785486 ADJ: 12.2998375493

Error report for Factored word to POS and morphology model, English to French translations, trained on limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 50.8359133127 NOM: 37.3362605196 PUN: 100.0 VER: 53.9632672789 KON: 22.9567744279 INT: 63.6363636364 PRO: 37.7546777547 DET: 15.0336700337 ABR: 100.0 NUM: 24.0 PRP: 39.4983991462 PRE: 63.6363636364 ADJ: 38.9185425853	ADV: 0.123839009288 NOM: 2.83937065496 VER: 16.9526341228 PRO: 13.0561330561 DET: 30.2861952862 PRP: 2.35859124867 ADJ: 13.1585054537

Error report for Factored word to POS and morphology model, English to French translations, trained on whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
ADV: 48.7306501548 NOM: 31.3208927918 PUN: 40.9448818898 VER: 52.319961334 KON: 22.8478023974 INT: 54.5454545455 PRO: 38.8773388773 DET: 15.2525252525 ABR: 100.0 NUM: 24.5 PRP: 40.8537886873 PRE: 63.6363636364 ADJ: 37.781387793	ADV: 1.57894736842 NOM: 2.70764727406 VER: 16.1188980184 KON: 2.21576462041 PRO: 14.6777546778 DET: 38.2828282828 PRP: 5.613660619 ADJ: 12.32304479

Error report for Moses baseline, French to English translations, trained on limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 56.400742115 VB: 60.0866824271 ": 100.0 RB: 55.3885630499 NN: 39.4619269619 FW: 68.4210526316 CC: 17.8272980501 WP: 51.4285714286 CD: 35.7843137255 TO: 36.9112814896 NNPS: 12.8342245989 PRP: 35.9340659341 JJ: 46.3657230298 IN: 41.2689877672 NNP: 15.1219512195 DT: 24.1579394282 VBinflec: 72.2235211597 UH: 50.0 EX: 60.7476635514	MD: 0 VB: 8.62884160757 NNPS: 4.8128342246 RB: 0 NN: 3.50621600622 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.357051772507 IN: 0 DT: 0 VBinflec: 18.0032733224 NNP: 0.889526542324

Error report for Moses baseline, French to English translations, trained on whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 51.2059369202 VB: 54.2947202522 ": 100.0 RB: 51.3929618768 NN: 35.0524475524 FW: 47.3684210526 CC: 17.938718663 WP: 48.0 CD: 31.3725490196 TO: 34.3921139102 NNPS: 11.7647058824 PRP: 33.2234432234 JJ: 42.3871461362 IN: 38.9837343729 NNP: 10.9325681492 DT: 24.9363147467 VBinflec: 65.9106850596 UH: 50.0 EX: 56.0747663551	MD: 0 VB: 9.14105594957 NNPS: 5.88235294118 RB: 0 NN: 3.51592851593 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.357051772507 IN: 0 DT: 0 VBinflec: 18.5410334347 NNP: 0.631276901004

Error report for Factored word to lemma models, French to English translations, trained on limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 57.6994434137 VB: 59.6926713948 ": 100.0 RB: 54.8387096774 NN: 39.4619269619 FW: 73.6842105263 CC: 18.6629526462 WP: 51.7142857143 CD: 34.8039215686 TO: 40.6900328587 NNPS: 13.9037433155 PRP: 35.2014652015 JJ: 46.5952563122 IN: 41.9142357844 NNP: 14.6628407461 DT: 25.4174922162 VBinflec: 72.504091653 UH: 50.0 EX: 61.6822429907	MD: 0 VB: 10.0078802206 NNPS: 3.20855614973 RB: 0 NN: 3.46736596737 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.382555470543 IN: 0 DT: 0 VBinflec: 17.979892448 NNP: 0.889526542324

Error report for Factored word to lemma models, French to English translations, trained on whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 51.3914656772 VB: 53.7431048069 ": 100.0 RB: 51.5029325513 NN: 34.7707847708 FW: 47.3684210526 CC: 17.9944289694 WP: 46.2857142857 CD: 30.3921568627 TO: 34.8849945235 NNPS: 11.2299465241 PRP: 33.5531135531 JJ: 42.0555980617 IN: 38.8358650356 NNP: 10.9612625538 DT: 24.9787715822 VBinflec: 66.3315407996 UH: 50.0 EX: 54.2056074766	MD: 0 VB: 9.33806146572 NNPS: 6.41711229947 RB: 0 NN: 3.58391608392 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.408059168579 IN: 0 DT: 0 VBinflec: 18.4708908113 NNP: 0.545193687231

Error report for Factored word to POS models, French to English translations, trained on limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 56.2152133581 VB: 59.6532702916 ": 100.0 RB: 55.1686217009 NN: 39.257964258 FW: 78.9473684211 CC: 18.4401114206 WP: 52.5714285714 CD: 35.7843137255 TO: 39.1018619934 NNPS: 13.3689839572 PRP: 35.4945054945 JJ: 46.49324152 IN: 40.8119370883 NNP: 14.7489239598 DT: 23.7758279083 VBinflec: 72.5274725275 UH: 50.0 EX: 61.6822429907	MD: 0 VB: 8.31363278172 NNPS: 5.34759358289 RB: 0 NN: 3.48679098679 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.357051772507 IN: 0 DT: 0 VBinflec: 17.6525602058 NNP: 0.918220946915

Error report for Factored word to POS models, French to English translations, trained on whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 51.2987012987 VB: 54.1765169425 ": 100.0 RB: 51.2463343109 NN: 34.6542346542 FW: 47.3684210526 CC: 18.2729805014 WP: 48.0 CD: 32.8431372549 TO: 35.8159912377 NNPS: 11.7647058824 PRP: 32.8205128205 JJ: 42.1321091558 IN: 38.9702917059 NNP: 10.9038737446 DT: 24.6249646193 VBinflec: 66.1444938041 UH: 50.0 EX: 57.0093457944	MD: 0 VB: 8.94405043341 NNPS: 5.88235294118 RB: 0 NN: 3.4188034188 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.331548074471 IN: 0 DT: 0 VBinflec: 18.3072246902 NNP: 0.631276901004

Error report for Factored word to POS and lemma models, French to English translations, trained on limited corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 55.4730983302 VB: 60.2048857368 ": 100.0 RB: 55.1686217009 NN: 38.9568764569 FW: 73.6842105263 CC: 18.6072423398 WP: 52.8571428571 CD: 38.7254901961 TO: 41.0186199343 NNPS: 13.9037433155 PRP: 35.9706959707 JJ: 46.1871971436 IN: 40.4355424116 NNP: 14.806312769 DT: 24.5825077838 VBinflec: 73.0652326397 UH: 50.0 EX: 61.6822429907	MD: 0 VB: 8.27423167849 NNPS: 4.8128342246 RB: 0 NN: 3.4965034965 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.357051772507 IN: 0 DT: 0 VBinflec: 17.1615618424 NNP: 0.889526542324

Error report for Factored word to POS and lemma models, French to English translations, trained on whole corpus

Mistranslated out of total words for POS	Mistranslated with right lemma but wrong inflection, out of total words for POS
MD: 51.6697588126 VB: 55.0433412136 ": 100.0 RB: 52.0894428152 NN: 34.7319347319 FW: 47.3684210526 CC: 18.6072423398 WP: 48.0 CD: 31.3725490196 TO: 36.7469879518 NNPS: 11.7647058824 PRP: 33.6996336996 JJ: 42.6931905126 IN: 39.1719317113 NNP: 11.018651363 DT: 25.247664874 VBinflec: 67.5707271452 UH: 50.0 EX: 55.1401869159	MD: 0 VB: 8.86524822695 NNPS: 6.41711229947 RB: 0 NN: 3.4188034188 CC: 0 WP: 0 CD: 0 TO: 0 PRP: 0 JJ: 0.331548074471 IN: 0 DT: 0 VBinflec: 17.9331306991 NNP: 0.545193687231